8-2021

Identification And Functional Characterization Of Plant Small Secreted Proteins During Arbuscular Mycorrhizal Symbiosis

Xiaoli Hu
xhu35@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Xiaoli Hu entitled "Identification And Functional Characterization Of Plant Small Secreted Proteins During Arbuscular Mycorrhizal Symbiosis." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Plant, Soil and Environmental Sciences.

Zong-Ming (Max) Cheng, Major Professor

We have read this dissertation and recommend its acceptance:

Xiaohan Yang, Robert N. Trigiano, Jessy L Labbé

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Identification And Functional Characterization Of Plant Small Secreted Proteins During Arbuscular Mycorrhizal Symbiosis

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Xiaoli Hu
August 2021

ACKNOWLEDGEMENT

During the past four years of my Ph.D. journey, I received tremendous help and encouragement from many people including my mentors, colleagues, friends, and family. First, I would like to thank my co-advisors, Dr. Zong-Ming (Max) Cheng and Dr. Xiaohan Yang. Thank you to Dr.Cheng who provided guidance throughout my studies and enriched my research activities by sponsoring my conference costs. Many thanks to Dr. Xiaohan Yang who worked closely with me on projects and assisted me in my science writing endeavors. Thank you to my committee members, Dr. Robert N. Trigiano and Dr. Jessy Labbe, who offered useful advice and suggestions which greatly enriched and improved my graduate career and dissertation. Special thanks to Dr. Trigiano who improved my English writing skills, and Dr. Labbe who provided his expertise in mycology in support of my fungal-related research. Furthermore, thank you to my clever and caring lab members Dr. Jin Zhang and Dr. Haiwei Lu who generously helped with my research as well as my daily life. Then, I would like to thank my friends and my family members who stayed with me and provided unwavering support for me, especially during such a tough period when the unprecedented pandemic impacted our work and life so much. It would be impossible for me to get my Ph.D. degree successfully without your love. Finally, I would like to thank China Scholarship Council which fully sponsored me to complete my graduate study.

ABSTRACT

Plant small secreted proteins (SSPs) are sequences of 50 – 250 amino acids in size which are transported out of cells to fulfill multiple functions related to plant growth and development and response to various stresses. With the development of more accurate and affordable genome sequencing technology, an increasing number of SSPs have been predicted using diverse computational tools based on machine learning. Although experimentally validated plant SSPs are still limited, some studies have reported that plant SSPs can be induced and involved in mutualistic relationships between plants and microbes. In Chapter I, known SSPs and their functions in various plant species are reviewed. Additionally, current computational tools and experimental methods that have been widely applied to identify plant SSPs are summarized. A new, robust, and integrated pipeline to discover plant SSPs is proposed. Furthermore, strategies for elucidating the biological functions of SSPs in plants are discussed in Chapter I. Chapter II presents predicted SSPs from 60 plant species and elucidates the evolutionary convergence of changes in SSP sequences. Furthermore, the expression of SSPs induced by arbuscular mycorrhizal fungi (AMF) which correspond to the convergent ability for different plants to form mutualistic association with AMF are explored. Overall, this study provides insightful ideas to understand functions of plant SSPs that occur during symbiosis between plants and fungi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ATTACHMENTS

INTRODUCTION

Plants face many microbes and insects throughout a life cycle. As the most important organ of plants, roots encounter diverse microorganisms within the rhizosphere including mycorrhizal fungi which form mycorrhiza. The term 'mycorrhiza' means 'fungus root', which is a symbiotic relationship between fungi and plant roots (Smith and Read 2010). This symbiosis is common in terrestrial ecosystems and nearly 95% of plant species form mycorrhizae characteristically which may have occurred primarily due to land colonization by plants (Read and Perez‐Moreno 2003; Smith and Read 2010). In mycorrhizal association, the fungi colonize the host plants' root tissues, either intracellularly as in arbuscular mycorrhizal fungi (AMF) or extracellularly as in ectomycorrhizal fungi (ECMF) (Johnson et al. 1997). Arbuscular mycorrhizal symbiosis (AMS) is the most ancient and broad type across mycorrhizal class of fungi. Ectomycorrhizal symbiosis is typically a beneficial interaction found in backwoods trees with ECMF (Bonfante and Genre 2010). Mycorrhizae impact the survival and wellness of plants and improves the plant microbial community structure. Additionally, mycorrhizae plays vital roles in plant water and nutrition uptake, such as phosphorus, from the soil to plants. In turn, plants make organic molecules such as sugars via photosynthesis and released to fungi via root exudates (Bolan 1991; Bonfante and Genre 2010). Furthermore, mycorrhizal fungi assist plants in developing resistance against the soil parasites, drought stress, and high concentrations of heavy metal. (Lehto 1992; Bellion et al. 2006).

Recently, studies on the mechanisms of mycorrhizae has gained much attention and improved our understanding of this association substantially (Shinano et al. 2011). For instance, Plett et al. (2014) unraveled how the effector protein MiSSP7 encoded by *Laccaria bicolor* initiates symbiosis with the host plant. MiSSP7 could affect the expression of jasmonic acid (JA) responsive genes by interacting with the host protein PtJAZ6, a negative regulator of JA-induced gene regulation in *Populus trichocarpa*. The association between MiSSP7 and PtJAZ6 protects PtJAZ6 from the JA-induced degradation. Furthermore, MiSSP7 blocks or mitigates the impact of JA on *L. bicolor* colonization of host roots. Vayssières et al. (2015) has shown that the *P. trichocarpa* - *L. bicolor* mutualistic association resulted in significant modifications in root architecture. Many short and swollen lateral roots formed and were sheathed by a fungal mantle, which affected the metabolism, signaling, and response to auxin in *P. trichocarpa* roots. The global analysis of auxin response gene expression and the regulation of auxin signaling F-BOX protein and auxin response factor expression in ECM roots indicates that symbiosis-dependent auxin signaling in root is activated during the colonization by *L. bicolor*.

In addition, various proteins secreted by plants can make a difference on carbon and nitrogen flow between soil and root interface, regulating both beneficial and harmful soil microbes (Jones et al. 2009; De-la-Pena et al. 2012). Proteins including peptidases, hydrolases and defensins have been revealed to affect plant-microbe mutualistic

interactions (De-la-Peña and M Loyola-Vargas 2012; Sagaram et al. 2013). Specifically, one typical category of proteins with 50 – 250 amino acids in size that can be transported out of cells is called small secreted proteins (SSPs). SSPs have been found to play important roles in various processes, including plant growth and development, plant response to abiotic and biotic stresses, and even beneficial plant–microbe interactions (Hu et al. 2021). For example, SSPs produced by legumes can enter the cytosol of nitrogen-fixing bacteria during nodule formation to govern the outcome of these mutualistic interactions (Wang et al. 2010a; Farkas et al. 2014). Cell-free recombinant peptide can enter the hyphae of *L. bicolor* and affect the growth of the fungus by a feeding experiment for several days (Plett et al. 2017). Whether a similar effect exists during the natural interaction between the host plant and fungus is still unknown.

To accurately identify plant SSPs and further understand their important functions, known SSPs and their functions including the maintenance of plant growth and development, response to abiotic and biotic stresses, and mediating mutualistic relationship between plants and microbes is summarized in Chapter I. Then, an update on the computational and experimental approaches that can be used to discover new SSPs is described. Finally, strategies for elucidating the biological functions of SSPs in plants is discussed. In Chapter II, SSPs are predicted from 60 diverse species to provide insight into the evolution of plant SSPs and identify plant SSPs that are highly related to symbiosis between plants and AMF. Here, whether the changes in sequence and gene expression of SSPs contribute to the evolutionary convergency of the ability to form symbiosis between plants and AMF is investigated. The observation from this study reveals such convergency shared among different plant lineages and several candidate SSPs which are highly related to AMS are identified.

Collectively, this work improves our knowledge of identification and functional validation of plant SSPs, especially the critical roles they play in symbiosis. Although SSP candidates were predicted, the pathways in which they are involved remain unclear. Overall, this study has laid a solid foundation for future research to understand functions of SSPs in symbiosis.

CHAPTER I

ADVANCES AND PERSPECTIVES IN DISCOVERY AND FUNCTIONAL ANALYSIS OF SMALL SECRETED PROTEINS IN PLANTS

This chapter contains one published manuscript
Hu, Xiao-Li, et al. "Advances and perspectives in discovery and functional analysis of small secreted proteins in plants." Horticulture Research 8.1 (2021): 1-14.
My contribution included: 1) writing introduction and conclusion, 2) writing main text of section 3 and 4, 3) creating figure 1, 4 and 5. Haiwei Lu was assisting with writing main text of section 2 and 5, she also created figure 2, 3 and 6.

Abstract

Small secreted proteins (SSPs) are less than 250 amino acids in length and are actively transported out of cells through conventional protein secretion pathways or unconventional protein secretion pathways. In plants, SSPs have been found to play important roles in various processes, including plant growth and development, plant response to abiotic and biotic stresses, and beneficial plant-microbe interactions. Over the past 10 years, substantial progress has been made in the identification and functional characterization of SSPs in several plant species relevant to agriculture, bioenergy, and horticulture. Yet, there are potentially a lot of SSPs that have not been discovered in plant genomes, which is largely due to limitations of existing computational algorithms. Recent advances in genomics, transcriptomics, and proteomics research, as well as the development of new computational algorithms based on machine learning, provide unprecedented capabilities for genome-wide discovery of novel SSPs in plants. In this review, we summarize known SSPs and their functions in various plant species. Then we provide an update on the computational and experimental approaches that can be used to discover new SSPs. Finally, we discuss strategies for elucidating the biological functions of SSPs in plants.

Introduction

Plant small secreted proteins (SSPs) are less than 250 amino acids (aa) in length and can be actively transported out of plant cells (Lease and Walker 2006; Plett et al. 2017). In plants, SSPs have been shown to play important roles in various biological processes such as growth, development, reproduction, resistance to abiotic and biotic stresses, and beneficial plant-microbe interactions (Chae and Lord 2011; Pan et al. 2012; Boschiero et al. 2020). In general, 30,000 – 40,000 protein-encoding genes have been reported in individual plant genomes (Sterck et al. 2007). Yet hundreds to thousands of SSPs are potentially overlooked in a single plant genome (Boschiero et al. 2019) for two reasons: 1) the SSP space is occupied by many proteins with a length of less than 100 aa (Nguyen et al. 2017; Plett et al. 2017), and 2) 50% of the discovered secreted proteins in plants do not have a known signal peptide (Krause et al. 2013), both of which create difficulties in SSP annotation using traditional computational approaches (Yang et al. 2011; Tavormina et al. 2015; Hellens et al. 2016).

In recent years, the increasing volume of genomics data and the continuously evolving machine learning algorithms have boosted the effectiveness of computationally predicting SSPs. Meanwhile, advances in functional genomics research have accelerated the experimental validation of predicted SSPs and the elucidation of their functional roles. As a result, SSP-focused research has become an emerging area with great potential for growth, as reflected by the rapidly increasing number of publications on SSPs in various organisms including animals, microbes, and plants. Here with a focus on plant SSPs, we first summarize the current understanding of SSP biosynthesis and secretion. We then discuss the structures and functions of representative SSPs that are well characterized in various plant species, including model species, food crops, bioenergy feedstocks, and horticultural plants. We also highlight computational tools, experimental approaches, and their combinations used to identify novel SSPs. Finally, we discuss the strategies that have been or can be used to explore the functions of SSPs.

Biosynthesis and secretion of SSPs in plants

Biosynthesis of SSPs in plants

In plants, SSPs have been found to be produced via multiple alternative pathways, as illustrated in Fig. 1.1. The majority of the characterized SSPs to date are proteolytic cleavage products synthesized via the removal of an N-terminal signal sequence (NSS; also known as N-terminal signal peptide) and/or a pro-domain from larger protein precursors, which can be either nonfunctional or functional (Tavormina et al. 2015; Chen et al. 2020). SSPs derived from nonfunctional precursors can be further classified into three subcategories based on features of their mature forms. SSPs belonging to the first subcategory typically consist of less than 20 aa in their mature forms which have few or no cysteine (Cys) residues and contain one to several types of post-translational modifications (PTM), such as tyrosine (Tyr) sulfation, proline (Pro) hydroxylation or Pro glycosylation. Therefore, these SSPs are named PTM SSPs. Several well-studied PTM SSPs in *Arabidopsis thaliana* are involved in plant growth and development, including CLAVATA 3 (CLV3), C-TERMINALLY ENCODED PEPTIDE 1 (CEP1), PLANT PEPTIDE CONTAINING SULFATED TYROSINE 1 (PSY1), and ROOT MERISTEM GROWTH FACTOR 1 (RGF1) (Murphy et al. 2012; Tabata and Sawa 2014; Tavormina et al. 2015). The second subcategory features SSPs with mature peptides that contain an even number (often ranging from 2 to 16) of Cys residues. These Cys residues are essential for forming the disulfide bonds in the active mature SSPs. Most of the known Cys-rich SSPs are involved in plant-microbe interactions, such as PLANT DEFENSINs (PDFs), nonspecific LIPID TRANSFER PROTEINS (nsLTPs), and KNOTTINs. Meanwhile, several Cys-rich SSPs have been found to regulate plant development, such as S-LOCUS CYSTEINE-RICH PROTEIN/S-LOCUS PROTEIN11 (SCR/SP11)

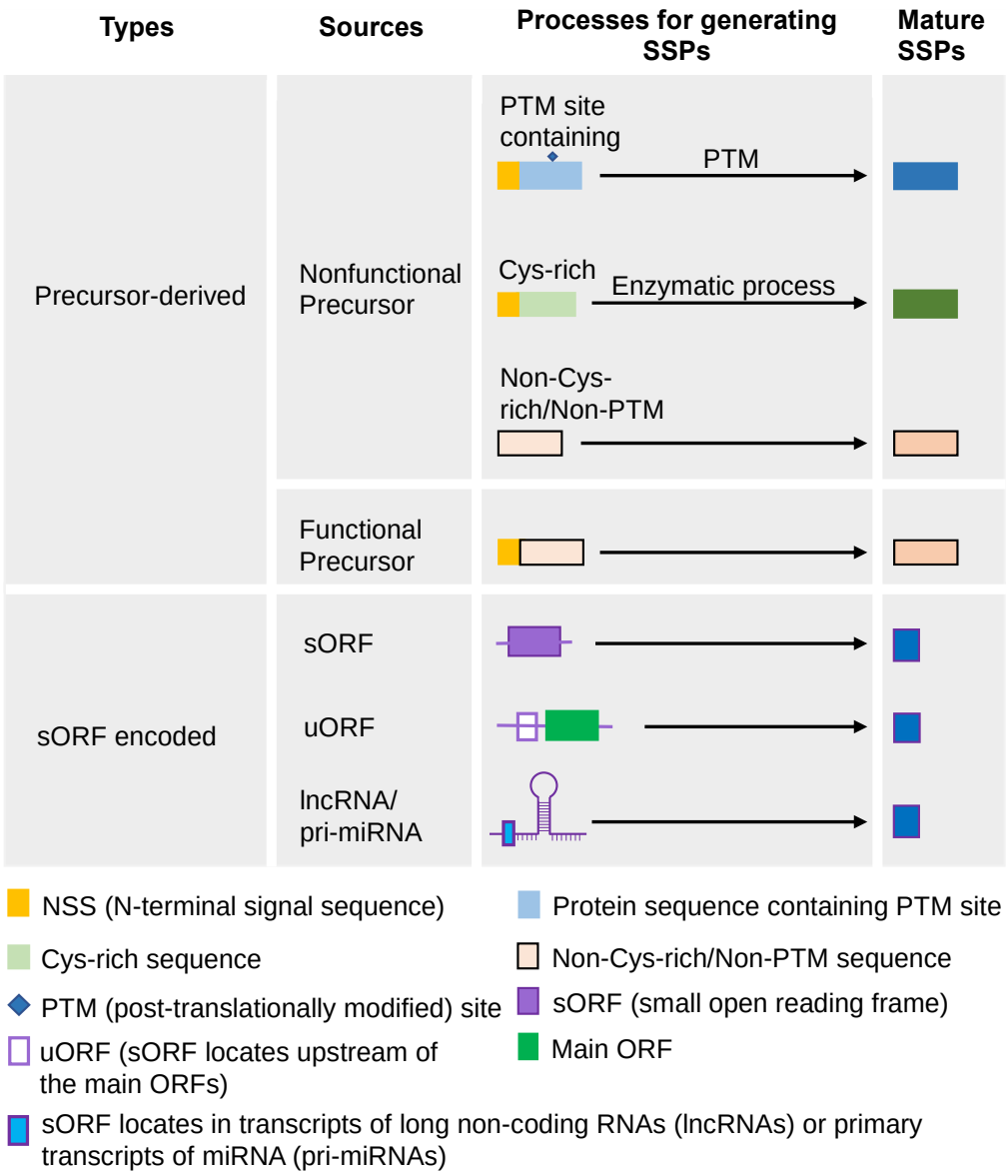

Figure 1.1 Classification of small secreted proteins (SSPs) in plants.

Plant SSPs can be derived from protein precursors, which can be either nonfunctional or functional, or translated from small open reading frames (sORFs). SSPs derived from functional protein precursors often contain an N-terminal signal sequence (NSS), which is removed during maturation. SSPs synthesized from nonfunctional precursors can be further divided into three categories: post-translationally modified (PTM) SSPs, Cs-rich SSPs, and non-Cys-rich/non-PTM SSPs. SSPs are encoded by sORFs that locate at upstream of main ORF (uORFs) and locate in transcripts of long non-coding RNAs (lncRNAs) or primary transcripts of miRNA (pri-miRNAs) and are transcripts encoding small proteins. Adapted from (Tavormina et al. 2015).

and LUREs (Tabata and Sawa 2014; Tavormina et al. 2015). The third subcategory contains non-Cys-rich/non-PTM SSPs, which often lack the NSS in their precursor forms and contain Cys, Pro, Tyr, glycine (Gly), lysine (Lys), or other amino acids with dominant roles in conferring the activity of the mature SSPs. SSPs within this subcategory have been primarily found to participate in plant defense responses, with SYSTEMINS (SYS), GRIM REAPER PEPTIDE (GRIp) and PLANT ELICITOR PEPTIDES (PEPs) being the representative examples (Tavormina et al. 2015). In the past decade, a growing number of plant SSPs has been found derived from functional protein precursors, such as INCEPTINs from *A. thaliana*, *Zea mays*, *Oryza sativa*, and *Vigna unguiculata*, the *Glycine max* SUBTILASE PEPTIDE (Gm-SUBPEP) and the *Solanum lycopersicum* CYSTEINE-RICH SECRETORY PROTEINS, ANTIGEN5, and PATHOGENESIS-RELATED 1 PROTEINS derived peptide 1 (CAPE1) (Tavormina et al. 2015).

In addition to being processed from larger protein precursors, plant SSPs can be directly encoded by small open reading frames (sORFs), which can sometimes locate upstream of the main ORFs (therefore called “uORFs”) or within presumed non-coding RNAs (e.g., long non-coding RNAs) or within primary transcripts of miRNAs. These SSPs are denoted as “short peptides encoded by sORFs”, “sPEPs”, or “nonprecursor-derived peptides” (Andrews and Rothnagel 2014; Tavormina et al. 2015; Hsu and Benfey 2018). Some known examples of such SSPs include the uORF2-encoded sucrose control peptide (SC-PEPTIDE) that is required for sufficient sucrose-induced repression of translation in *A. thaliana* (Rahmani et al. 2009), the miPEP171b that regulates root development in *Medicago truncatula* (Lauressergues et al. 2015) and ENOD40s that are involved in sucrose use in nitrogen-fixing nodules in *G. max* (Röhrig et al. 2002).

Plant SSPs can be directly translated from small open reading frames (sORFs) or derived from protein precursors, which can be either nonfunctional or functional. SSPs derived from protein precursors often contain an N-terminal signal sequence (NSS), which is removed during peptide maturation. SSPs synthesized from nonfunctional precursors can be further divided into three categories: post-translationally modified (PTM) SSPs, Cys-rich SSPs, and non-Cys-rich/non-PTM SSPs. uORF: sORFs located upstream of the main ORFs. Adapted from (Tavormina et al. 2015).

Mechanisms of SSP secretion in plants

Our knowledge of plant SSP secretion largely overlaps with our understanding of protein trafficking and secretion, which follows several different mechanisms (Ding et al. 2014; Goring and Di Sansebastiano 2018; Wang et al. 2018b). The majority of plant SSPs with an NSS are secreted via the conventional protein secretion (CPS) pathway (Fig. 1.2) conserved among eukaryotes. Guided by their NSS, SSPs are first transported to the endoplasmic reticulum (ER) where the NSS is removed. These SSPs

are then exported to the cis side of the Golgi apparatus (Golgi) and further sorted through the Golgi or the trans-Golgi network (TGN). Modifications, such as glycosylation, that are required for SSPs maturation occur when SSPs travel through the Golgi. Finally, the mature SSPs are delivered to the apoplast via secretory vesicles or granules (Goring and Di Sansebastiano 2018; Hsu and Benfey 2018; Wang et al. 2018b; Zhang et al. 2019a).

However, some NSS-containing SSPs bypass the CPS pathway and follow unconventional protein secretion (UPS) routes (Fig. 1.2) (Goring and Di Sansebastiano 2018; Wang et al. 2018b) traveling to the extracellular space, usually upon pathogen attack or the exposure to other biotic or abiotic stress conditions (Krause et al. 2013; Zhang et al. 2019a). The simplest UPS route directly transports these proteins from the ER to the plasma membrane (PM). Alternative UPS routes utilize vesicular carriers, including the secretory multivesicular body (MVB) and vacuole, that can fuse with the PM to release their contents into the apoplast/extracellular space (Goring and Di Sansebastiano 2018).

In addition, secreted proteins without an NSS (also known as cytosolic leaderless proteins, LSPs), which represent a large proportion of the plant secretome (Ding et al. 2014), cannot be processed by the CPS. These proteins have been proposed to be secreted through the excyst-positive organelle (EXPO) – a double-membrane organelle whose formation is Golgi- and TGN-independent and can fuse with the PM to secrete LSPs (Fig. 1.2) (Krause et al. 2013; Ding et al. 2014).

Known SSPs and their biological roles in plants

Known SSPs

Because the genome of model herbaceous plant *A. thaliana* is considered to be better annotated and characterized than other plant species, we focus on known SSP families found in *A. thaliana*. Also, we discuss SSPs that have been identified from several important plant species, including *Z. mays*, *O. sativa*, *S. lycopersicum*, *M. truncatula* and *P. trichocarpa*. A large number of SSPs have been computationally predicted in plants, as demonstrated in public databases including OrysPSSP (Pan et al. 2012), PlantSSP (Ghorbani et al. 2015), and MtSSPdb (Boschiero et al. 2020). For instance, according to the database PlantSSP (Ghorbani et al. 2015), there are 2,451, 5,373, and 3,216 predicted SSPs, which are less than 200 aa in length with NSS, in *A. thaliana*, O. sativa and *P. trichocarpa*, respectively. These predicted SSPs account for 6.9%, 8.0%, and 7.1% of all the annotated proteins (including splice variants) in the *A. thaliana* (version TAIR10), O. sativa (version MSU6.1), and *P. trichocarpa* (JGI v2) genome, respectively. More recently, with the reannotation of the *M. truncatula* genome, 4,439 genes (6.3% of all the annotated genes) were predicted to encode SSPs that are less

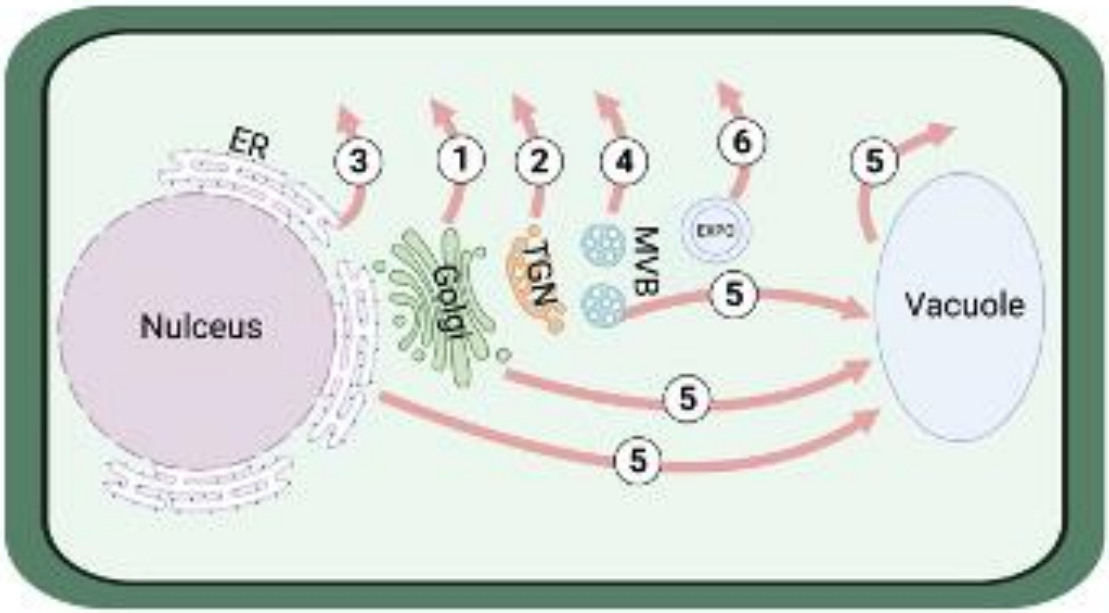

Figure 1.2 Various secretion mechanisms of small secreted proteins (SSPs) in plants.

Most N-terminal signal sequence (NSS)-containing SSPs are secreted via the conventional protein secretion (CPS) pathway which starts at the endoplasmic reticulum (ER). Then the SSPs rout through the Golgi apparatus (Golgi) (1) and sometimes trans-Golgi network (TGN) (2) before being delivered to apoplast. Meanwhile, some NSS-containing SSPs are secreted via various unconventional protein secretion (UPS) routes, including direct transportation from ER to apoplast (3), and the employment of secretory multivesicular body (MVB) (4) and vacuole (5). Cytosolic leaderless proteins (LSPs) are secreted through the excyst-positive organelle (EXPO) (6). Adapted from (Goring and Di Sansebastiano 2018).

than 230 aa with NSS but not transmembrane regions (Boschiero et al. 2020). Although interest in decoding genomes for potential SSPs has been growing substantially in recent years, only a limited number of SSPs have been experimentally characterized, which are distributed among approx. 50 gene families (Chen et al. 2020), with their representative members listed in Table A1.

Structure of known SSPs in plants

Protein function is dependent on a well-defined and folded three-dimensional (3D) structure and intrinsically disordered regions (IDRs), which are not likely to form a defined 3D structure (van der Lee et al. 2014). Some of the known SSPs in plants have well-defined 3D structure, as demonstrated in Fig. 1.3. For instance, hydroxyproline-bound tri-arabinoside induced conformation was found when post-translationally modified protein CLV3 became biologically active (Shinohara and Matsubayashi 2013). The β-turn-like conformation, for example, which is a feature of CEP1, is associated with biological activity (Bobay et al. 2013). On the other hand, enzymatic maturation processes produce bioactive Cys-rich SSPs with correct oxidative folding under oxidative conditions by forming diverse disulfide patterns as well as loop regions, which are supposed to be crucial for protein-protein interactions (Moroder et al. 2005; Tabata and Sawa 2014). SCR/SP11 contains an α/β sandwich motif connected by L1 loop that serve as binding site for specific receptors (Mishima et al. 2003). LTP has four α-helices, three loops and four disulfide bridges with eight conserved cysteines (Chae and Lord 2011). EPF includes one loop and three disulfide bonds, which contains two antiparallel β-strands connected by a 14-residue loop (Ohki et al. 2011). However, it has been estimated that 10% of secreted proteins are intrinsically disordered proteins (IDPs), with >70% of their length being IDRs (van der Lee et al. 2014). For example, LTP1 from *A. thaliana* contains a defined 3D structural domain (Fig. 1.3C) and without IDR (Fig. 1.4A) but LEA4 from *A. thaliana* has no defined 3D structural domain and is fully disordered (Fig. 1.4B).

Biological roles of known plant SSPs

Role of SSPs in plant growth and development

Some of the known SSPs are associated with multiple aspects of plant growth and development. During these processes, most SSPs act as signaling molecules that are involved in cell-to-cell communication by binding membrane receptors and coordinating responses with plant hormones (Murphy et al. 2012; Fukuda and Ohashi-Ito 2019). In terms of meristem maintenance, CLE14 and CLE40 expression has been observed in *A. thaliana* root meristematic zone and observed to play roles in controlling meristematic

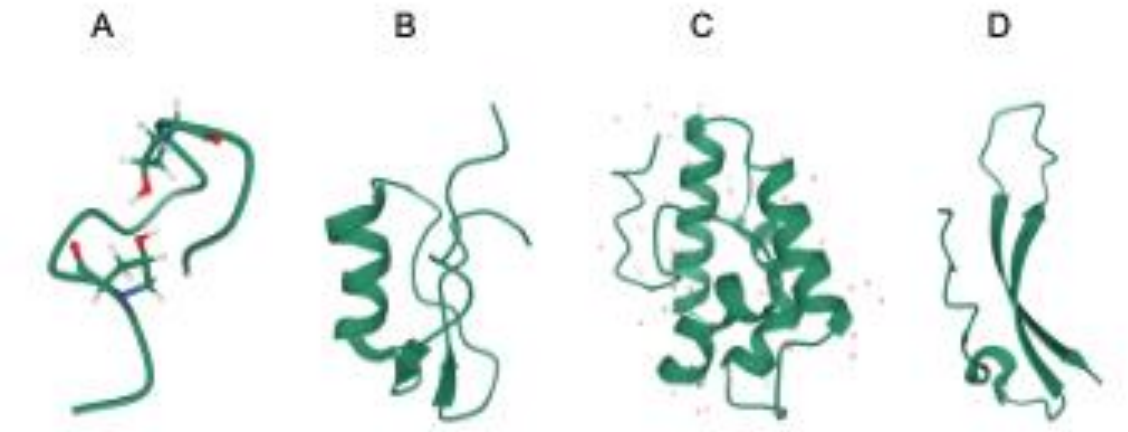

Figure 1.3 Three-dimensional structure of some known small secreted proteins in plants.

(A) CEP1 (PDB ID: 2MFO). (B) SCR/SP11 (PDB ID:1UGL). (C) LTP (PDB ID:1MZL). (D) Stomagen (PDB ID: 2LIY). The Protein Data Bank (PDB) data were obtained from RCSB protein data bank (https://www.rcsb.org/) (Berman et al. 2000; Burley et al. 2020) and visualized using Mol* (Sehnal et al. 2018).

activity as well as cell number (De Smet et al. 2008; Meng and Feldman 2010). Although CLE43 does not affect root apical meristem (RAM) growth in *A. thaliana* (Whitford et al. 2008), its homologues, BnCLE43a and BnCLE43b, were found in Brassica napus could repress *A. thaliana* root growth when synthetic peptides were added to the culture medium (Han et al. 2020). In *A. thaliana*, both CLE9 and CLE10 control xylem differentiation through regulation of the cytokinin signal pathway (Fukuda and Hardtke 2020), and CLE41 can drive vascular cell division (Etchells and Turner 2010). In contrast, PtrCLE20 identified in vascular cambium cells of *P. trichocarpa* was shown to restrain cell division, resulting in an inhibition of lateral growth of the stem (Zhu et al. 2020). Besides the impact on vegetative tissues or organs, SSPs can affect flower development. For example, CLV1 acts with CLV3 to avoid enlarged meristems and extra floral organs in *A. thaliana* (Fletcher et al. 1999). The pollen-specific SlPRALF gene that encodes a 129 aa preproprotein was recognized to negatively regulate pollen tube elongation in *S. lycopersicum* (Covey et al. 2010).

Role of SSPs in plant response to abiotic and biotic stresses

To sense and respond to various stresses, plants have evolved complex signaling and defense mechanisms (Chagas et al. 2018). Induced SSPs have been observed in many stress responses in plants, including some SSPs recognized as hormone-like molecules (Segonzac and Monaghan 2019). SSPs act quickly and synergistically at low concentrations in reaction to different stresses (Wang and Irving 2011).

SSPs are involved in a variety of biotic stresses responses in diverse plant species. For example, an SSP called SYSTEMIN identified in *S. lycopersicum* was the first wound response signaling peptide (Pearce et al. 1991; Constabel et al. 1998). When plants are attacked by herbivores or pathogens, a series of defense signals and pathways can be induced by SYSTEMIN through its interaction with SYSTEMIN RECEPTOR 1, which includes stimulation of PROTEASE INHIBITOR production, as well as enhancement of ethylene and jasmonic acid biosynthesis (Kandoth et al. 2007; Wang et al. 2018a). Plant SSPs can initiate immune responses and increase resistance to pathogens. For example, an SSP called IRP, which was identified from the proteomic analysis of *O. sativa* suspension cells cultured with bacterial peptidoglycan and fungal chitin, increased the abundance of phenylalanine ammonia-lyase 1 (PAL1) and activated mitogen-activated protein kinases (MAPKs), which are known to be associated with plant immunity (Wang et al. 2020). Two pathogen-responsive SSPs, TaSSP6 and TaSSP7, are responsible for resistance to Septoria tritici blotch, a severe foliar disease caused by the fungal pathogen *Zymoseptoria tritici* in *Triticum aestivum* (Zhou et al. 2020). In *Z. mays*, Zip1 was demonstrated to trigger plant immunity by activating salicylic acid defense signaling (Ziemann et al. 2018).

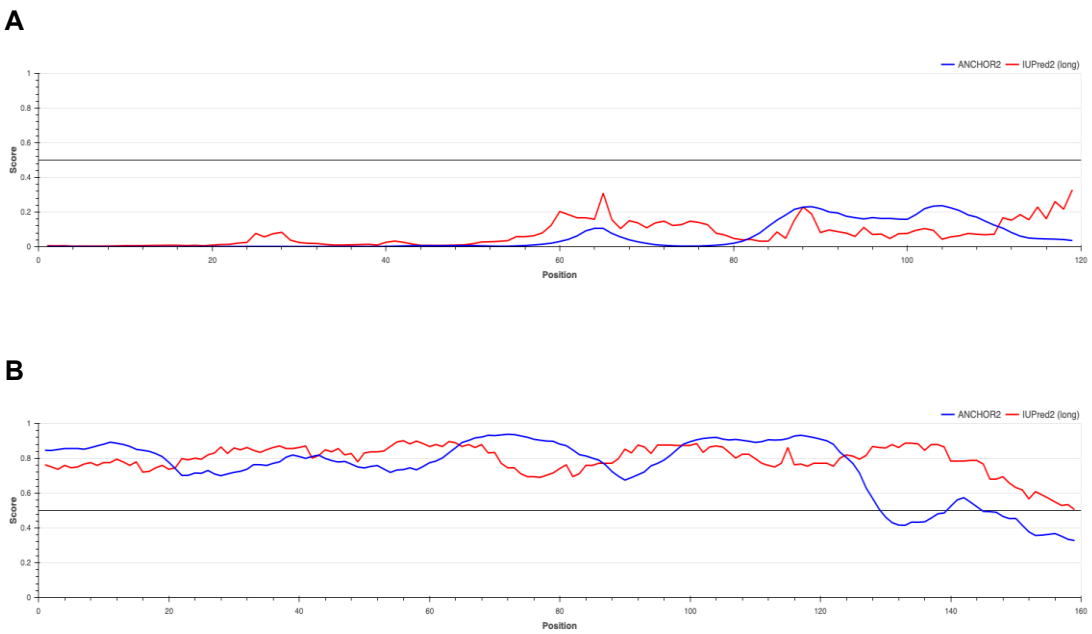

Figure 1.4 Examples of plant small secreted proteins containing intrinsically disordered regions (IDRs).

(**A**) LTP1 (gene locus: AT2G38540), with a defined 3D structural domain (PDB ID: 1MZL). (**B**) LEA4 (gene locus: AT5G06760) with IDR only. The protein sequence data were obtained from Phytozome (https://phytozome-next.jgi.doe.gov/) and IDRs were predicted using IUPred2A (https://iupred2a.elte.hu/).

SSPs are also involved in responses to abiotic stresses. For example, CLE25, found in A. *thaliana*, is induced under dehydration, which triggers ABA biosynthesis in leaves to prevent water loss by regulating stomatal closure (Takahashi et al. 2018). In *A. thaliana* roots, AtRALFL8 encoding an SSP can be induced not only by nematode infection, but also by drought stress, leading to cell wall remodeling (Atkinson et al. 2013). To determine extracellular proteins that respond to heat stress, a quantitative proteomic analysis was conducted by collecting proteins from heat tolerant *Sorghum bicolor* cell suspension culture medium, resulting in the identification of an SSP named germin protein, which was highly induced at the protein level (Ngcala et al. 2020). Another example is the small peptide AtPep3 encoded by *AtPROPEP3* which has been shown to play an important role in salinity stress tolerance in *A. thaliana* (Nakaminami et al. 2018).

Role of plant SSPs in beneficial plant-microbe interactions

SSPs play important roles in cross-kingdom interactions. It is widely accepted that SSPs generated from plant-associated microorganisms (e.g., fungi, bacteria) can be used as effector proteins to promote plant microbial colonization (Stergiopoulos and Wit 2009; Kohler et al. 2015; Trivedi et al. 2020). However, studies on the identification of plant SSPs as effector proteins that affect microbes have been very limited (Plett et al. 2017). Plants can adapt to a low availability of nutrients by altering root system architecture, with some can form symbiotic associations with rhizobia and mycorrhizal fungi (Péret et al. 2009; Imin et al. 2013). In legumes, SSPs can affect root development and rhizobial-legume symbiosis (Gonzalez-Rizzo et al. 2006; Whitford et al. 2012). CLE family members have been characterized in different species, such as CLE12 and CLE13 in *M. truncatula*, CLE-RS (CLE-root signal) 1/2/3 in Lotus japonicus and RIC (rhizobium-induced CLE) in *G. max*. These SSPs appear to be involved in the negative systemic autoregulation of the nodulation (AON) pathway and inhibit newly formed nodules in roots (Laffont et al. 2020). Conversely, in *M. truncatula*, CEP1 was found to modulate lateral root formation and increase the number and size of nodules (Imin et al. 2013). When *L. japonicus* was inoculated with the arbuscular mycorrhizal (AM) fungus *Rhizophagus irregularis*, in comparison with formation of nodules in L. japonicus, alternate CLE genes, including LjCLE19 and LjCLE20, were upregulated in roots, indicating that different signaling pathways are involved in arbuscular mycorrhizal and root nodule symbiosis (Handa et al. 2015). In addition, a recent study reported that SSPs produced by *P. trichocarpa* were induced when co-culture with ectomycorrhizal mycorrhizal (EM) fungus *L. bicolor* and several *P. trichocarpa* SSPs could enter fungal hyphae when they were exposed to *L. bicolor* (Plett et al. 2017), suggesting plant SSPs may mediate ectomycorrhizal symbiosis as well.

Computational and experimental approaches for discovery of SSPs in plants

Computational approaches for discovery of SSPs

In general, there are two main steps to computationally predict SSPs in plant genomes, i.e., predicting small proteins encoded by sORFs and subsequently evaluating their ability to be secreted. A large number of sORFs can be found by locating in-frame start and stop codons in the plant genomes. However, annotations of sORFs have been largely overlooked because such short sequences were initially classified as random nonsense occurrences (Martinez et al. 2020). In the recent decade, progress in the development of computational methods for gene prediction has contributed to the identification of numerous sORFs in plants. For example, sORF finder is a tool for identifying putative small sORFs between 10 and 100 amino acids based on significant selective constraints, which works well for predicting sORFs in plant genomes (Hanada et al. 2010). Small Peptide Alignment Discovery Application (SPADA) is a homology-based program which can accurately identify and annotate genes in a given family, including sORFs in plants (Zhou et al. 2013). One caveat of these in silico sORF prediction tools is that the predicted sORFs may be pseudogenes. To address this issue, transcript expression data generated by transcriptome sequencing (RNA-seq) can be used for identifying functional sORFs, as demonstrated in SSP discovery in P. trichocarpa (Yang et al. 2011; Plett et al. 2017). Transcript sequences obtained from RNA-seq data can be either protein coding sequences (CDS) or non-coding RNAs (Liu et al. 2017; Mewalal et al. 2019). Finally, using DeepCPP, a new deep neural network based tool, aims to predict short sequences with coding potential (Zhang et al. 2020).

The potential for secretion of small proteins has been determined using several alternate tools based on specific algorithms, in particular many use newly developed machine learning (ML) approaches (Table A2). To predict NSS-containing SSPs, SignalP 5.0, based on deep neural networks, is commonly utilized because it has a user-friendly interface and good performance across plant species (Almagro Armenteros et al. 2019). However, since an NSS is common in several types of membrane proteins, membrane spanning proteins with both predicted signal peptide and at least one transmembrane region should be excluded (Uhlén et al. 2015). MEMSAT-SVM (Nugent and Jones 2012) can be used for transmembrane helix topology prediction, and SPOCTOPUS (Viklund et al. 2008) is designed for predicting both signal peptide and transmembrane topology. Because the existence of certain numbers of NSS-containing proteins follow UPS routes, SecretomeP has been constructed and is a ML algorithm to predict unconventionally secreted proteins (Nielsen et al. 2019). In addition, the number of Cys residues, and their arrangement, have been used to predict Cys-rich SSPs without signal peptide (Li et al. 2014). In some studies, an additional criterion, such as the lack of endoplasmic reticulum-retention motif, is taken into consideration for secretion prediction. Several authors recommend

that small proteins containing C-terminal KDEL or HDEL motifs should be excluded as non-SSPs (Li et al. 2014; de Bang et al. 2017). Protein secretion mediated by conventional (e.g., CLE (Whitewoods 2021)) or unconventional (e.g., PME (Wang et al. 2016)) mechanisms can be evaluated using various tools for predicting multiple protein subcellular localizations, such as LocTree3 (Goldberg et al. 2012; Goldberg et al. 2014), CELLO (Yu et al. 2006), YLoc (Briesemeister et al. 2010), DeepLoc (Almagro Armenteros et al. 2017), and TargetP (Armenteros et al. 2019a). Also, ML-based methods have been developed recently for predicting both conventional and unconventional secretion, e.g., ApoplastP (Sperschneider et al. 2018), BUSCA (Savojardo et al. 2018) and Plant-mSubP (Sahu et al. 2020) . A pipeline integrating the best methods for computational prediction of SSPs is proposed in Section 4.3.

Experimental approaches for discovery of SSPs

The putative SSPs predicted using computational approaches described in Section 4.1 need to be verified using experimental approaches to provide protein-level evidence. To address this issue, protein mass spectrometry (MS) data can be used to determine 1) whether the predicted SSPs are truly expressed proteins in extracellular localization and 2) whether the predicted SSP sequences are full-length or partial fragments of longer protein sequences. For instance, a novel 15 aa secreted peptide named CEP1 encoded by AT1G47485 was effectively identified in *A. thaliana* by liquid chromatography-mass spectrometry (LC-MS) analysis (Ohyama et al. 2008a). The feasibility of this system was tested initially by detecting a known small secreted peptide CLE44 in the medium using transgenic *A. thaliana* overexpressing the CLE44 gene. Computational prediction of SSP secretion can also be verified through MS analysis of extracellular proteins. For example, protein MS has been successfully used to identify plant immune response proteins that are secreted into apoplastic space in *A. thaliana* leaves (Rutter and Innes 2017). Proteomic analyses of secretomes have identified secreted proteins in *O. sativa* (Shinano et al. 2011), *Hippophae rhamnoides* (Gupta and Deswal 2012), *S. bicolor* (Ngcala et al. 2020), *Solanum chacoense* (Liu et al. 2015), and *S. lycopersicum* (Briceño et al. 2012). Such global analyses of plant secretomes could facilitate the discovery of SSPs. However, proteins containing IDRs of sufficient length tend to be more susceptible to degradation, resulting in lower protein abundance (van der Lee et al. 2014). This may cause a problem for studying plant SSPs that contain a large portion of IDRs using proteomics approaches because MS has lower sensitivity than transcriptome sequencing. To increase the sensitivity of detecting SSPs in plants, it is necessary to enrich for IDRs containing proteins and low-molecular weight proteins in protein extract using gel-filters (Chen et al. 2015) or ultrafiltration devices (Greening and Simpson 2010; Villalobos Solis et al. 2020).

Besides plant secretome proteomics, molecular approaches can be used to test SSP secretion. For example, the CDS of SSPs can be fused with reporter genes, such as green fluorescent protein (GFP) (Zhang et al. 2017), and the gene fusion constructs can

be tested for secretion of reporter-tagged SSPs using agroinfiltration-based transient gene expression (Norkunas et al. 2018) or stable transformation in plants. The secretion of SSPs has been tested using the yeast expression system as well (Plett et al. 2017).

Integrative approaches for discovery of SSPs

From an amalgamation perspective, multiple tools can be assimilated to predict SSPs. Here we propose such a pipeline for SSP discovery by integrating the methods discussed above (as illustrated in (Fig. 1.5). Briefly, sORFs encoding small proteins are predicted from genomic sequences using gene prediction pipeline such as Seqping (Chan et al. 2017) based on self-training HMM models and transcriptomic data. Next, NSS-containing small proteins that are transported via CSP pathways are predicted with ML based tools, such as SignalP 5.0. At this stage small proteins containing transmembrane regions, which are unlikely to be secreted, should be identified and eliminated from downstream analysis. Given that some NSS-containing proteins follow USP pathways, additional ML-based software, such as SecretomeP, may be applied simultaneously. In addition, the secretion ability of proteins without an NSS are inferred by subcellular localization prediction tools (Table A2), which are helpful for predicting secreted proteins contaning an NSS as well. Putative SSPs predicted by computational tools are then validated with MS-based and/or molecular experiments, particularly for their secretion ability, before further functional characterization. Proteomics data is then used to confirm the protein expression of putative sORFs, to discover small proteins that are derived from larger protein precursors and/or to localize protein accumulation outside cells.

Strategies for elucidating the function of plant SSPs

Examination of the secretion and transport pathways

Given that apoplastic localization of SSPs can be vital for their function, functional characterization of SSPs often requires refining the knowledge of their trafficking, transport, and secretion routes both within plants and between plants and their microbial partners. Perhaps the most direct method for investigating SSP movement is to visualize SSPs under a fluorescence or electron microscope after tagging them with a fluorescent protein or other label, as demonstrated by Wang et al. (2010b) when investigating EXPO-mediated transportation of the *A. thaliana* Exo70 paralog – Exo70E2, and by Chen et al. (2016) when studying the movement of the transcription factor HY5 from shoot to root in *A. thaliana*. One requirement for this approach is that the fusion of the SSPs and the fluorescent markers must not alter the mobility, secretion, or the function of the SSPs (Wang et al. 2018b; Burko et al. 2020) or interfere with the folding and fluorescence intensity of the markers.

Small molecule reagents have been used to dissect protein trafficking routes. A widely used example is the fungal toxin brefeldin A (BFA). Given that BFA can disrupt the retrograde traffic from the Golgi to the ER, it serves as a powerful tool for distinguishing Golgi-dependent and -independent protein trafficking (Zhang et al. 2011; Pinedo et al. 2012). Another example is concanamycin A (ConcA) – an inhibitor of vacuolar-type ATPase (V-ATPase), which blocks post-Golgi trafficking and has been used in examining the transportation pathway of VHA-a3 (Scheuring et al. 2011; Viotti et al. 2013). Additionally, small molecules that can interact with trafficking-related organelles or vesicles have been used to screen for their potential application in elucidating protein secretion pathways (Rodriguez-Furlan et al. 2018). The power of these trafficking inhibitors, however, becomes limited when it comes to examining the movement of SSPs between plants and microbes. An alternative approach could be based on fluorescently tagged SSP, which was discussed above and appears to be more useful for examining the cross-kingdom movement of plant SSPs.

In addition, a learn-by-design approach based on rewriting the transport pathway can be informative for evaluating if secretion is required for SSP function. Targeted redirection has been achieved by fusing SSPs to alternative sorting signals. For example, Rojo et al. (2002) fused different vacuolar sorting signals (VSSs) to the C-terminus of CLV3 and redirected the destination of CLV3 from apoplast to the vacuole. The authors concluded that apoplastic localization is essential for CLV3 to activate the CLV signaling pathway in *A. thaliana*.

Uncovering phenotypic traits conferred by SSP-encoding genes

Reverse genetics techniques, by imparting loss- or gain-of-function mutations via ectopic expression, virus-induced gene silencing (VIGS), and RNA interference (RNAi) (Gilchrist and Haughn 2010; Ben-Amar et al. 2016), are among the most powerful tools to reveal phenotypes associated with genes of interest. These techniques work equally well for studying the function of SSP-encoding genes. For example, CLV3 – the meristem development regulator, when constitutively overexpressed in transgenic *A. thaliana* (Brand et al. 2000) demonstrated the correlation between the level of CLV3 protein and the accumulation of the meristem cells. In addition, *A. thaliana* in which the expression of CLV3 was suppressed by RNAi was created by Chuang and Meyerowitz (2000) for studying the associated phenotypic changes in floral development. Similarly, RNAi-induced suppression of the PtCLV3 ortholog PttCLE47 were employed by Kucukoglu et al. (2020) to investigate its role in cambial development and secondary xylem formation in hybrid aspen (*P. tremula* × *P. tremuloides*).

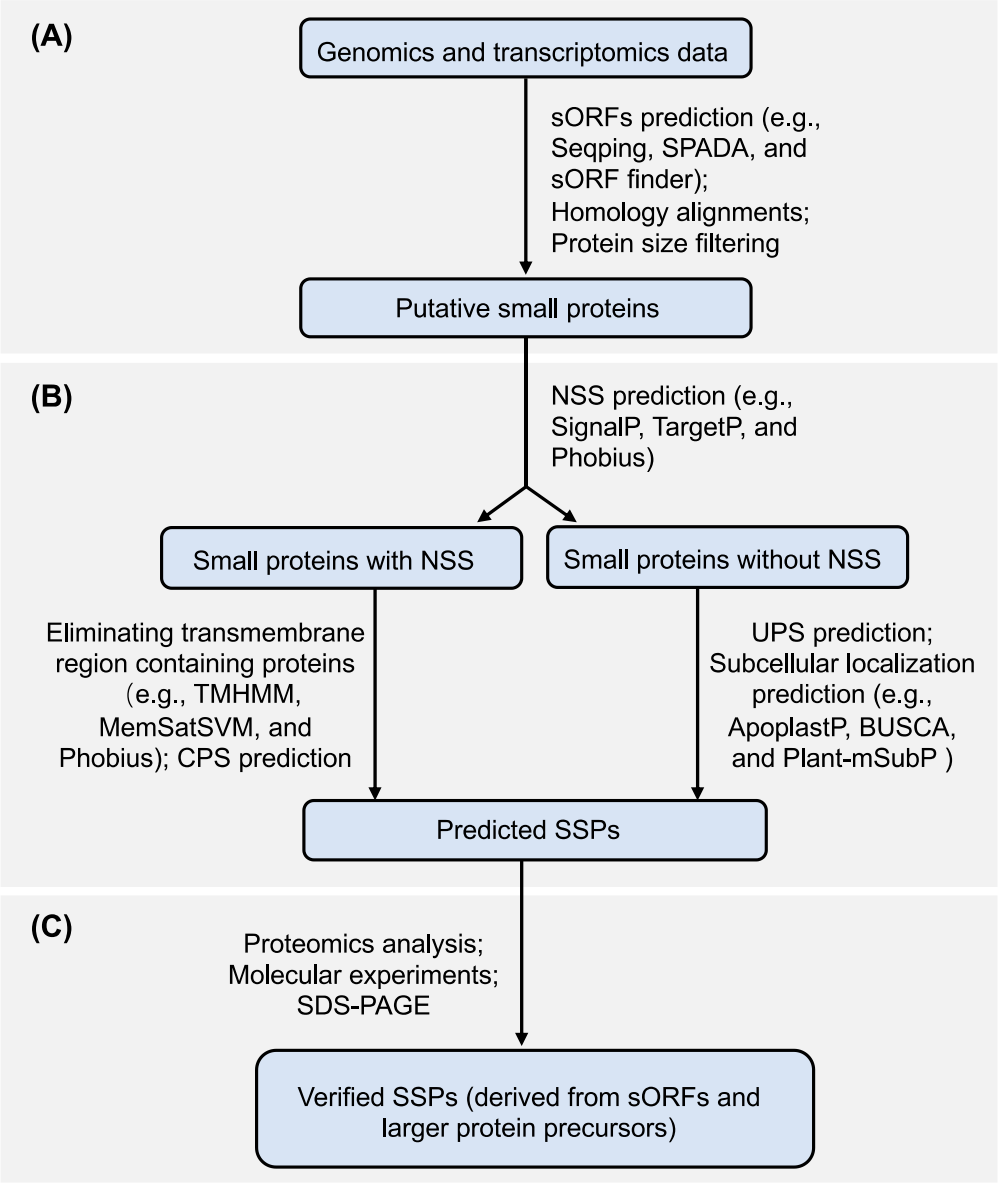

Figure 1.5 An integrative pipeline for discovery of small secreted proteins (SSPs) in plants.

(**A**) Small open reading frames (sORFs) encoding small proteins can be predicted by using gene prediction tools based on genome sequence and transcriptomic data. (**B**) Predicting secretion processes for small proteins using machine learning approaches. (**C**) Experimental validation of predicted SSPs. NSS: N-terminal signal sequence for protein secretion; CPS: conventional protein secretion; UPS: unconventional protein secretion; MS: mass spectrometry; SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis.

Besides traditional techniques, the recent revolution in gene editing tools, particularly the invention of the CRISPR/Cas and related technologies, provides new opportunities for efficient gene knock-out, gene knock-in, gene activation, and gene suppression in plants (Liu et al. 2016; Hassan et al. 2020; Yang et al. 2020; Zhang and Qi 2020). Its development is based on an immune system naturally found in bacteria and archaea, the CRISPR/Cas9 system has been widely used for creating gene knockouts by creating double-strand breaks (DSBs), which are then repaired by error-prone the non-homologous end joining (NHEJ) in plants and therefore often lead to indel mutations in the target gene. The efficacy of CRISPR/Cas9-mediated gene knockout has been demonstrated in a number of herbaceous and woody plant species (Xue et al. 2015; Elorriaga et al. 2018; Li et al. 2019; Liu et al. 2019b). In the last few years, the adaptation of CRISPR into a recruiting platform and the discover of Cas9 variants have made CRISPR/Cas a more versatile tool. For example, transcriptional activation and suppression of single and multiple genes can now be conferred by the CRISPR/deactivated Cas9 (dCas9) based transcriptional regulation system (Lowder et al. 2017; Zhang et al. 2019b). All of these tools can be used in tuning the expression of SSPs for revealing their targets and examining their biological impacts.

Identification of receptors and partners involved in SSP signal transduction pathways

As discussed above, many plant SSPs act as signaling molecules and have the ability to affect the expression of other genes. Therefore, identifying the receptors and other downstream targets of an SSP of interest is the ultimate step towards deciphering SSPs' biological function. A number of early studies, particularly those done in *A. thaliana*, have been relying on creating targeted mutants or performing mutational screen to achieve this goal. Taking receptors of CLV3 in *A. thaliana* for instance: CLV1, which is a leucine-rich repeat (LRR) receptor-like kinase (RLK), was verified via phenotypic analysis of single or double mutants (Clark et al. 1995). Meanwhile, CORYNE (CRN) which is a membrane associated protein kinase, and TOADSTOOL2 (TOAD2) which is a receptor-like kinase, were identified by screening the population created with ethyl methanesulfonate (EMS) mutagenesis (Müller et al. 2008; Kinoshita et al. 2010).

Besides mutational screens, protein-protein interaction (PPI) data can provide valuable evidence in identifying novel partners that interact with SSPs during signal transduction. Several in vitro and in vivo PPI detection approaches, such as affinity purification (AP), tandem affinity purification (TAP) and yeast two-hybrid (Y2H), have been commonly used (Rao et al. 2014). In particular, the capability of Y2H-based approaches has been extended from one-by-one clonal identification to proteome-wide mapping of PPIs, with the recent development of matrix-based Y2H methods coupled with next-generation sequencing (NGS) technology (Erffelinck et al. 2018). Compared with mutational

screen, Y2H-NGS approaches make it possible to identify novel interaction partners of SSPs even within an organism whose genome has not been fully annotated yet.

Discovery-based extraction, screening, and identification of SSPs

High-throughput analytical approaches that couple selective enrichment, fractionation/isolation, and phenotype screening followed by MS-based identification provide an established framework to screen plant tissues for biologically relevant SSPs (Pearce et al. 1991; Pearce et al. 2001; Ohyama et al. 2008a; Cao et al. 2019; Demarque et al. 2020) (Fig. 1.6). This classical approach for the discovery of novel natural products starts with an enrichment strategy to selectively isolate molecules of interest from highly complex crude extracts. For SSPs, common cellular extraction techniques use size exclusion ultrafiltration strategies, such as molecular weight cut-off (MWCO) spin column filters, to selectively enrich for low molecular weight protein fractions (Greening and Simpson 2010; Villalobos Solis et al. 2020). Other techniques include gel-based separations (Cheli and Baldi 2011; Chen et al. 2015; Wang et al. 2020), solvent extractions (Ohyama et al. 2008a; Patel et al. 2018), and size exclusion chromatography (Mohd-Radzman et al. 2015; Patel et al. 2018). Following these enrichment strategies, SSPs can be further fractionated based on physicochemical properties (e.g., polarity, hydrophobicity, stability, solubility) using liquid chromatography (Alexandersson et al. 2013; Kim et al. 2013; Wilson et al. 2020).

Either as crude extract mixtures, enrichments, or isolated fractions, SSPs can be evaluated for their bioactivity against cell-based or cell-free biosystems. Cell-based screening can be used to assess simple effects on cell viability, morphology, and proliferation, or to elucidate the mechanism of action. Common phenotypes profiled in cell-based systems are growth promotion/restriction or antimicrobial activity (Matsubayashi and Sakagami 1996; Ito et al. 2006; Runyoro et al. 2006; Mabona et al. 2013). Alternatively, cell-free screening has been employed to evaluate the effect of SSPs to better describe the thermodynamic, kinetic or structural basis for molecular interactions with other cellular constituents (Makarewich and Olson 2017). Cell-free screening can be employed to identify SSPs with the abilities to scavenge free radicals, chelate metals, or bind to certain macromolecular targets that regulate various biological processes such as epigenetic processes and cell proliferation (Nwachukwu and Aluko 2019; Ding et al. 2020).

Following the detection of fractions with relevant bioactivity, molecule libraries can be further interrogated via high-throughput LC-MS/MS to sequence unknown SSPs. Some of the current challenges in accurate and sensitive identification of SSPs with MS include lack of SSP representation in protein databases, inadequate understanding of SSP maturation mechanisms, and partial knowledge of their post-translational modifications. Thus, the characterization of SSPs by LC-MS/MS can benefit from the use of *de novo* search strategies (Cheng et al. 2010). *De novo* sequencing algorithms

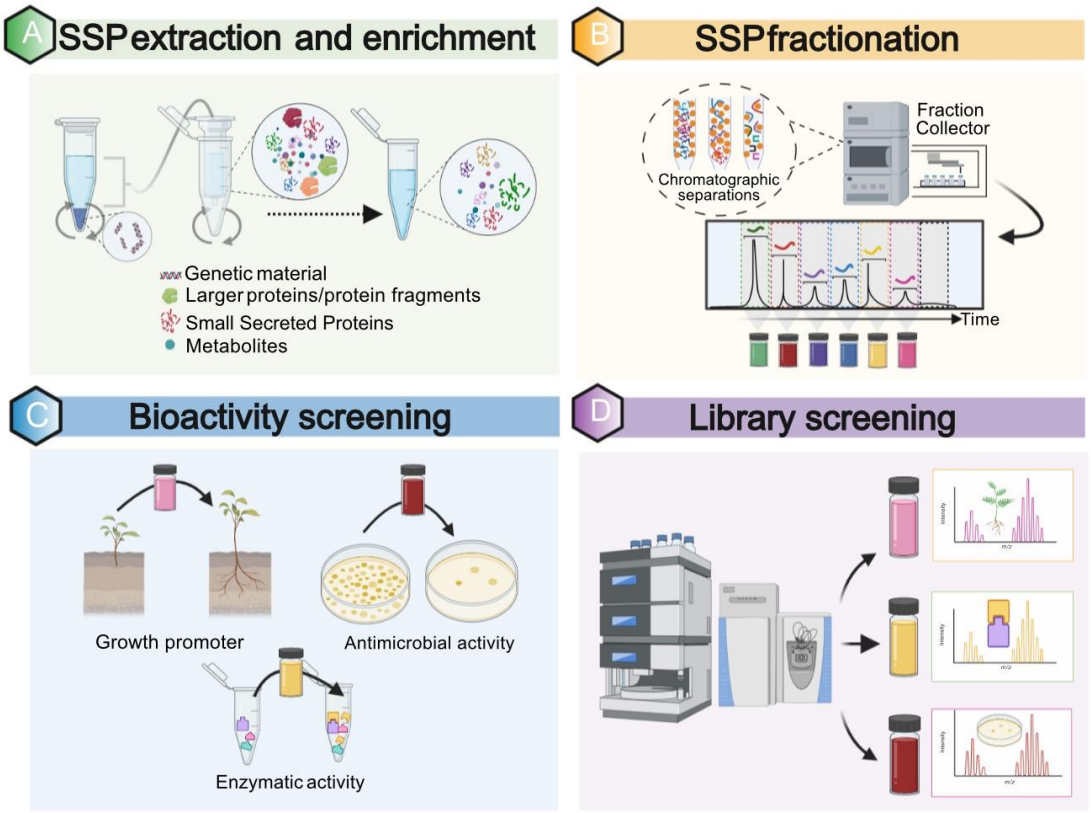

Figure 1.6 Experimental framework to screen biologically relevant small secreted proteins (SSPs).

The experimental workflow to characterize bioactive SSPs consists of four main steps: (A) The extraction and enrichment of the low molecular weight (MW) fraction of the secreted proteome of a sample e.g., with the use of molecular weight cut-off filters. (B) The fractionation/isolation of low MW fractions using different chromatographic separations techniques to reduce their complexity and assemble a set of SSP candidates to test for bioactivity. Other low MW molecules like metabolites can be removed at this step if needed. (C) SSPs bioactivity assays against cell-based or cell-free systems to elucidate their mechanisms of action (i.e., growth promotion or antimicrobial activity). (D) Interrogation of SSP fraction libraries with bioactivity via high-resolution/high-mass accuracy LC-MS/MS. Novel SSP sequence characterization could be aided by *de novo* search strategies. Figure was created with BioRender.com.

derive peptide sequences using only fragment ion information from the tandem mass spectra, are generally optimized to run without the restriction of cleavage enzymes (i.e., trypsin) and work in an unbiased manner as they do not necessarily require any input based on prior knowledge of the sample (Ma and Johnson 2012).

Conclusion and perspectives

In the past several years, there has been increasing evidence that SSPs play important roles during plant growth, development, and response to biotic and abiotic stresses, and consequently a growing appreciation of the biological significance of plant SSPs. A sheer number of SSPs have been predicted in diverse lineages of organisms, and the intercellular or inter-organismal movement of SSPs infers that SSPs are likely a significant and common mode of signaling among organisms. It is now known that SSPs are synthesized and secreted via diverse pathways in plants. Currently, however, the number of characterized SSPs in plants is low. The majority of SSPs encoded in plant genomes are overlooked and remain unannotated. Roadblocks that prevent progress in the study of SSPs include 1) a lack of reliable methods for isolating SSPs for experimental characterization, 2) a lack of capabilities for real-time monitoring the intercellular or inter-organismal movement of SSPs, 3) a lack of structural data for SSPs, and 4) a lack of computational tools for predicting non-conventional secretion of SSPs.

Recent advances in high-throughput molecular screening approaches and bioinformatics offer exciting opportunities for the discovery and characterization of SSPs. For example, the rapid accumulation of omics data, including genomics, transcriptomics and proteomics, provide rich databases for discovering plant SSPs, including those derived from larger protein precursors and directly encoded by sORFs. Meanwhile, advanced ML tools have evolved to predict the secretion pathways, including both CPS and UPS that SSPs follow. Such computational prediction on secretion can be verified experimentally, for example, via bioimaging of fluorescent reporter tagged protein candidates. In addition, advanced plant biotechnologies, particularly, CRISPR/Cas-based genome-editing systems and transcriptional regulation systems (i.e., CRISPRa and CRISPRi) allow for efficient gene knock-out, activation, and suppression, and therefore analysis of the biological roles of SSPs, and identification of their partners by combining with PPI and NGS data. The discovery and functional role of SSPs in plant growth and development will continue to expand in the near future.

CHAPTER II

PHYLOGENOMIC ANALYSIS OF PLANT SMALL SECRETED PROTEINS ASSOCIATED WITH ARBUSCULAR MYCORRHIZAL SYMBIOSIS

This chapter will be submitted for publication.
Xiao-Li H., …… 2021. Phylogenomic analysis of plant small secreted proteins associated with arbuscular mycorrhizal symbiosis.

Xiaoli Hu will be considered as first author. My contribution included: 1) conceiving study, 2) collecting data and data analysis, 3) writing manuscript and creating figures. Jin Zhang assisted with RNA-Seq analysis and creating figures.

Abstract

Arbuscular mycorrhizal symbiosis (AMS) is an ancient and widespread mutualistic association between plants and fungi, which plays an essential role in nutrient exchange, enhancement in plant stress resistance, development of host, and sustainability of ecosystem. To date, an increasing volume of studies have shown that plant small secreted proteins (SSPs) are involved in multiple biological processes, such as plant growth and development, response to abiotic and biotic stresses, and beneficial symbiotic interactions. In this study, we performed computational prediction of SSPs in 60 plant species including 39 AMS species and 21 non-AMS species. Through comparative genomics analysis, we identified two types of ortholog groups containing SSP genes: (i) AMS-specific ortholog groups containing SSPs only from at least 30% of the AMS species in this study and (ii) AMS-preferential ortholog groups containing SSPs from both AMS and non-AMS species, with AMS species containing significantly more SSPs than non-AMS species. Also, we analyzed gene expression in four AMS species and one non-AMS species and identified plant SSP genes responsive to arbuscular mycorrhizal fungus (AMF) *R. irregularis*. Furthermore, we examined the diversification and conservation in 3D protein structure and promoter regions between genes in the AMS-preferential ortholog groups containing AMF-inducible SSPs. Finally, we identified genes co-expressed with the *P. trichocarpa* SSP genes in the AMS-preferential ortholog groups through co-expression network analysis. Our results provide new insights into the molecular basis of AMS evolution as well as expand our understanding of the function of plant SSPs during AMS.

Introduction

Plant small secreted proteins (SSPs) are usually less than 250 amino acids (aa) in length, which are derived from large precursor or encoded by small open reading frames (sORFs) (Lease and Walker 2006; Tabata and Sawa 2014; Tavormina et al. 2015). SSPs play roles in many biological processes, such as plant growth and development, response to various stresses, and mediation of intercellular communications (Fukuda and Ohashi-Ito 2019; Chen et al. 2020). For instance, in *A. thaliana*, CLE3 is involved in the regulation of lateral root formation (Araya et al. 2014).

Root-derived CLE25 transmits water deficiency signals to leaves through vascular tissues in *A. thaliana* and therefore improves dehydration tolerance (Takahashi et al. 2018). In *M. truncatula*, overexpression of CEP1 leads to inhibition of lateral root development and enhancement of nodulation (Mohd-Radzman et al. 2015). A total of 417 putative plant SSPs have been identified to be significantly regulated during the process of forming mutualistic symbiosis between *P. trichocarpa* roots and the ectomycorrhizal fungus *L. bicolor*, indicating that plant-derived SSPs play potential roles in cross-kingdom interactions (Plett et al. 2017).

Discovering SSPs in plants can be started from mining sORFs in the sequenced plant genomes. With the affordability of genome sequencing and recent advances in transcriptomics, high-throughput identification of sORFs is getting much easier. Based on two commonly used metrics, sequence conservation and sequence similarity (Peeters and Menschaert 2020), multiple bioinformatics methods have been developed to aid the prediction of sORFs, such as sORF finder, which is an evolutionary selective constraints-based tool (Hanada et al. 2010), and SPADA, which is a sequence homology-based software (Zhou et al. 2013). Furthermore, various tools have emerged for assessing the coding potential of putative sORFs, such as Coding-Non-Coding Identifying Tool (CNIT) based on support vector machine (SVM) (Guo et al. 2019), MiPepid based on logistic regression model (Zhu and Gribskov 2019), and DeepCPP based on deep neural network (Zhang et al. 2020). After generating sORF candidates, machine learning based methods can be used for secretion prediction. Prediction of conventional secretion is primarily achieved by predicting N-terminal signal peptides through SignalP (Armenteros et al. 2019b) and excluding proteins containing transmembrane regions, which can be predicted by TMHMM (Möller et al. 2001). In addition, unconventional secretion of proteins that do not have N-terminal signal peptides can be predicted by SecretomeP (Nielsen et al. 2019) , ApoplastP (Sperschneider et al. 2018), BUSCA (Savojardo et al. 2018), Plant-mSubP (Sahu et al. 2020), etc.

Different pipelines that combine several methods have been used for SSP prediction. For example, a list of predicted novel SSPs in *M. truncatula* was created by using multiple sequential filtering steps, including protein length selection (<230 aa), signal peptide identification, and removal of proteins containing transmembrane helices and endoplasmic reticulum-retention signals (de Bang et al. 2017). In another study, discovery of SSPs in *P. trichocarpa* based on RNA-Seq datasets was achieved by selecting complete ORFs that encode proteins of less than 250 aa in length, followed by prediction of protein secretion using three different tools (Plett et al. 2017) .

Arbuscular mycorrhizal symbiosis (AMS) is one of the most ancient and broadly occurring mutualistic associations between plants and arbuscular mycorrhizal fungi (AMF) (MacLean et al. 2017). This intimate relationship mainly improves plant mineral nutrition acquisition, which potentially enhances crop yield (Hu et al. 2021). In addition, it would increase plant tolerance to biotic and abiotic stresses (Yang et al. 2014; Bona

et al. 2017; Lanfranco et al. 2018; Hu et al. 2021). AMS also contributes to many ecosystem functions, including soil aggregation, less fertilizer utilization, and reduction of nutrient losses (Rillig et al. 2019).

Over the last two decades, based on the alteration of symbiosis phenotypes in gene knockout or knockdown mutants, a lot of genes have been identified to be involved in AMS (MacLean et al. 2017). Recently, with the availability of rich plant genomic resources, phylogenomics provided great opportunity for studying evolutionary pattern of conserved genes in plants in relation to AMS (Delaux 2017). Recently, the expression of two SSP genes *LjCLE19* and *LjCLE20* in *Lotus japonicus* was regulated by AMF *R. irregularis* (Handa et al. 2015). More recently, some putative sORF-encoding genes in *Populus* were reported to be responsive to *R. irregularis* (Mewalal et al. 2019).

The goal of this study is to gain a better understanding of the relationship between plant SSPs and AMS. To achieve this goal, we predicted SSPs in 60 sequenced plant genomes using a computational pipeline and identified candidate plant SSP genes that are potentially involved in AMS through phylogenetic analysis of ortholog groups containing SSP genes and identification of gene expression responsive to AMF. Furthermore, we performed comparative analysis of 3D protein structure and the promoter regions between genes in selected ortholog groups which were either specific to or predominately represented by AMS plant species. Finally, we built co-expression networks *P. trichocarpa* genes to identify other genes associated with the *P. trichocarpa* SSP genes in the ortholog groups predominately represented by AMS plant species. Our results indicate that convergency in SSP sequences and gene expression induced by fungi is related to convergent emergency of AMS in diverse plant species. The SSP candidates identified in this study lay a valuable foundation for experimental characterization of AMS-related genes to gain deep understanding of the molecular mechanisms underlying the interactions between plants and AMF.

Materials and Methods

Plant species and protein sequences

Primary protein sequences (i.e., the longest protein sequence for each gene) were downloaded from Phytozome13 (https://phytozome-next.jgi.doe.gov) for a total of 60 plant species representing diverse plant lineages (Table S1). Symbiosis status of the plant species was determined based on the published literatures (Wang and Qiu 2006; Brundrett 2009).

Construction of ortholog groups and phylogenetic trees

The primary protein sequences of the 60 species were used as input to construct ortholog groups using Orthofinder (Emms and Kelly 2019). For constructing gene trees, the protein sequences of each mostly single-copy orthologue group, which contains no more than 3 genes in each plant species, aligned using MAFFT version7 (Katoh et al. 2019). The protein sequence alignments were further trimmed by removing sites with more than 50% gaps or Ns and removing sequences less than 50% of the alignment in length. The trimmed protein sequence alignments were used to create gene trees using the maximum likelihood approach implemented in IQ-Tree 2 (Minh et al. 2020) (default parameters; 1000 bootstrap replications), with the best-fitting substitution models determined by ModelFinder (Kalyaanamoorthy et al. 2017). Then the species tree was generated from the gene trees by performing coalescent-based analysis using ASTRAL (Mirarab et al. 2014).

Prediction of SSPs

We created a computational pipeline to predict SSPs from a total of 1,911,840 protein sequences in 60 plant genomes (Fig. 2.1). Briefly, small proteins (encoded by complete ORFs both start and stop codons) of 50-250 amino acids in length were selected as an initial small protein subset. The secretion prediction for proteins in the initial small protein subset was performed using eight either widely used or recently released methods based on different algorithms. Specifically, SignalP 5.0 (Armenteros et al. 2019b), Phobius (Käll et al. 2007), and TargetP (Armenteros et al. 2019a) were used for the prediction of N-terminal signal sequence (NSS). TMHMM 2.0 (Möller et al. 2001), MEMSAT-SVM (Nugent and Jones 2012), and Phobius (Käll et al. 2007) were used for the prediction of membrane domains. ApoplastP (Sperschneider et al. 2018) , DeepLoc (Almagro Armenteros et al. 2017), and Plant-mSubP (Sahu et al. 2020) were used for the prediction of protein subcellular locations. Stand-alone applications of these selected methods were run on a computer cluster. The principle of majority-decision called MDSEC as previously described (Uhlén et al. 2015) was used to predict SSPs (i.e., small proteins containing NSS predicted by at least two out of the three approaches, including SignalP 5.0 (Armenteros et al. 2019b), Phobius (Käll et al. 2007), and TargetP (Armenteros et al. 2019a), were considered to be secreted proteins). As NSS can also be found in membrane proteins, small proteins containing at least one transmembrane region predicted by each single tool were eliminated from the pool of predicted NSS-containing SSPs, resulting in the first list of predicted NSS-containing SSPs without transmembrane regions. In addition, a great number of proteins without NSS can be secreted via unconventional secreted pathway (Hu et al. 2021). Thus, we generated the second list of SSPs with extracellular location predicted by two out of the three approaches including ApoplastP (Sperschneider et al. 2018), DeepLoc (Almagro Armenteros et al. 2017), and Plant-mSubP (Sahu et al. 2020). Finally, a set of non-redundant predicted SSPs were generated by merging the first and the second lists of

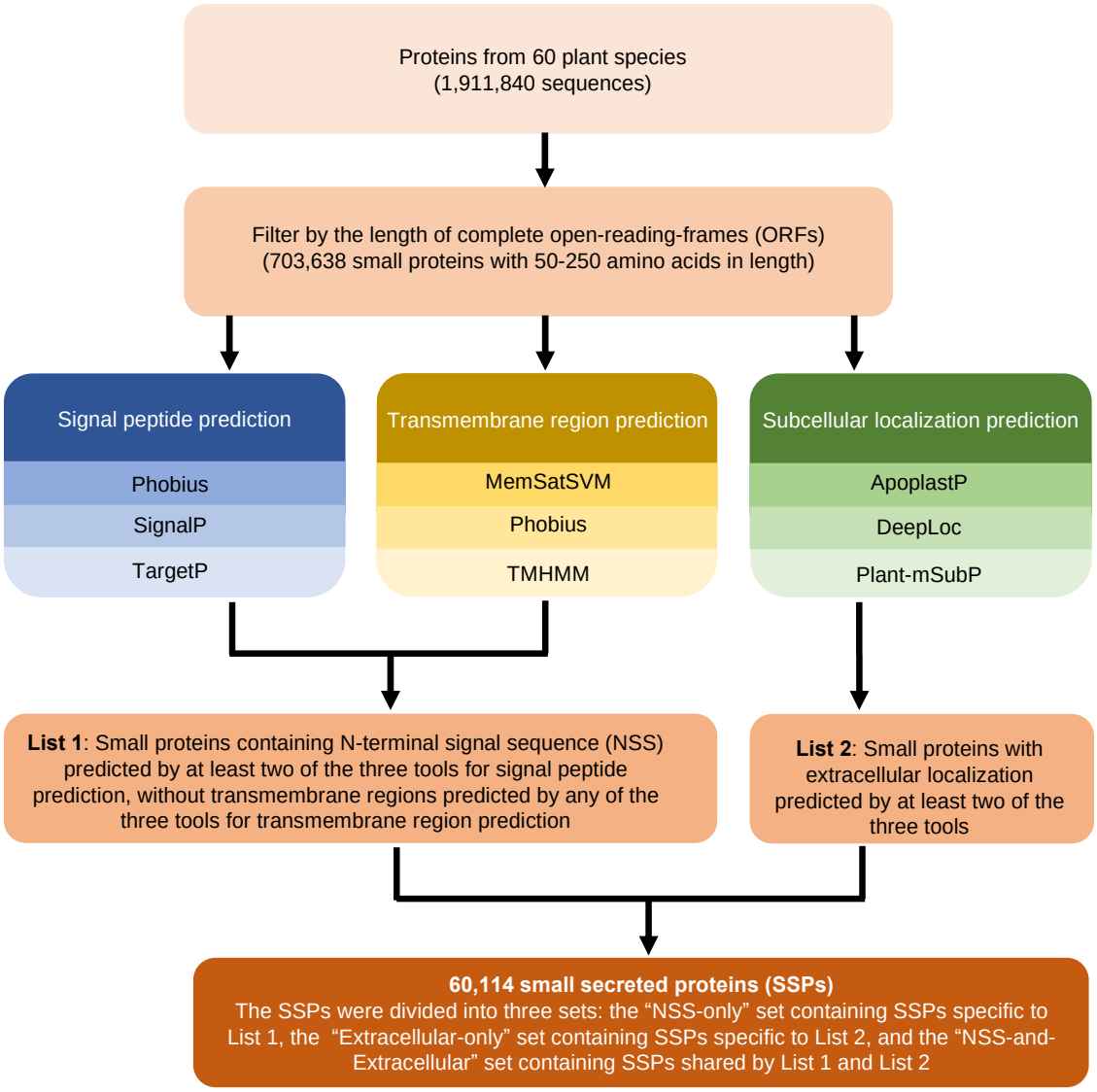

Figure 2.1 A computational pipeline used for predicting small secreted proteins (SSPs) in plant genomes.

The input was primary protein sequences of 60 plant species listed in Table S1. Small proteins with a full-length of 50-250 aa were identified for secretion prediction using different methods. Conventional protein secretion featured by N-terminal signal sequence (NSS) were predicted by using SignalP 5.0 (Armenteros et al. 2019b), Phobius (Käll et al. 2007), and TargetP (Armenteros et al. 2019a). Transmembrane domains were identified by using TMHMM 2.0 (Möller et al. 2001), MEMSAT-SVM (Nugent and Jones 2012), and Phobius (Käll et al. 2007). Extracellular protein localization was predicted by using ApoplastP (Sperschneider et al. 2018), DeepLoc (Almagro Armenteros et al. 2017), and Plant-mSubP (Sahu et al. 2020).

predicted SSPs mentioned above, which were further divided into three sub-categories: NSS-only (from the first list only), NSS-plus-extracellular (shared by both the first and the second lists), Extracellular-only (from the second list only).

RNA-Seq data analysis

We performed a cross-species comparative transcriptome analysis using public RNA-Seq data of different plant roots inoculated with AMF, which include four AMS species, including *Cucumis sativus, Manihot esculenta, M. truncatula* and *T. aestivum*, as well as one NAMS species *A. thaliana* as a control (Table S2).

The raw reads retrieved from the National Center for Biotechnology Information Sequence Read Achieve (http://www.ncbi.nlm.nih.gov/sra/) were filtered with the BBDuk program from JGI's BBTools (https://jgi.doe.gov/data-and-tools/bbtools) to trim adapters and extremities with a quality value per base lower than 20. After trimming adapter sequences and filtering out low-quality reads, the clean reads were mapped to the latest genome assembly for each species using STAR2.7.9a (Dobin et al. 2013). The mRNA abundance of each gene in each species was quantified as FPKM. Differentially expressed genes (DEGs) in each species were determined by applying EBSeq (Leng et al. 2013) in the R package. The cut-off for significant DEGs were absolute log2(fold change) >1 and false discovery rate (FDR) corrected *P* value < 0.05.

Promoter analysis

The promoter sequences (2 kb upstream of translation initiation site) of representative upregulated and non-upregulated SSPs in different species were downloaded from Phytozome (https://phytozome-next.jgi.doe.gov). Conserved parts for AMS inducible gene of SSPs in promoter regions were analyzed using online server PlantPAN 3.0 (Chow et al. 2019) with default parameter.

Protein structural modeling

The 3D structures of SSPs and their closely related proteins in the AMS-preferential ortholog groups were predicted using the Phyre2 web portal (Kelley et al. 2015). The protein structural alignments were constructed and visualized using PyMol (https://pymol.org/2/).

Co-expression network analysis

For co-expression network construction, the expression data was obtained in the *Populus* Gene Atlas Study from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). Pearson correlation coefficients (PCCs) were calculated in parallel between all pairs of gene expression vectors. A threshold of *P* value < 0.05 and absolute PCC $\geqslant$ 0.95 were applied to identify the significant correlations, and their co-expression relationships were visualized by Cytoscape (Shannon et al. 2003). Functional classification of the co-expressed genes of candidate SSPs was carried out with MapMan (Thimm et al. 2004).

Results

Identification of SSPs in 60 plant species

From the 60 plant species listed in Table S1, we predicted two lists of SSPs using the computational pipeline illustrated in Fig. 2.1. The first SSP list included 23,360 SSPs containing N-terminal signal sequence (NSS), without transmembrane regions (Figs. S1a and S1b). The second SSP list contained 48,081 SSPs with extracellular localization predicted by at least two methods (Fig. S1c). By combining these two SSP lists, we generated a non-redundant list of 60,114 SSPs (Table S3), which were divided into three sets: (i) the NSS-only set containing 12,033 SSPs from the first SSP list only, (ii) the Extracellular-only set containing 36,754 SSPs from the second SSP list only, and (iii) the NSS-and-extracellular set containing11,327 SSPs shared by the two SSP lists (Fig. S1d). The distribution of SSP numbers in each plant species was illustrated in Fig. 2.2.

AMS-related ortholog groups

We identified 60,981 ortholog groups accounting for 91.6% of total number of protein sequences from 60 plant species listed in Table S1. Among these, 9,390 ortholog groups contain 49,472 predicted SSPs, which account for 82.3% of total number of SSPs predicted from the 60 plant species ortholog group. The SSP-containing ortholog groups were divided into three types: 6,629 AMS-specific ortholog groups contained SSPs from AMS plant species only, 1,817 ortholog groups contained SSPs from non-AMS plant species only, and 944 ortholog groups contained SSPs from both AMS and non-AMS species. Aiming to identify ortholog groups that are highly associated with AMS status, we firstly selected three AMS-specific ortholog groups containing proteins from at least 30% of the 39 AMS species, including OG0000442 (containing genes

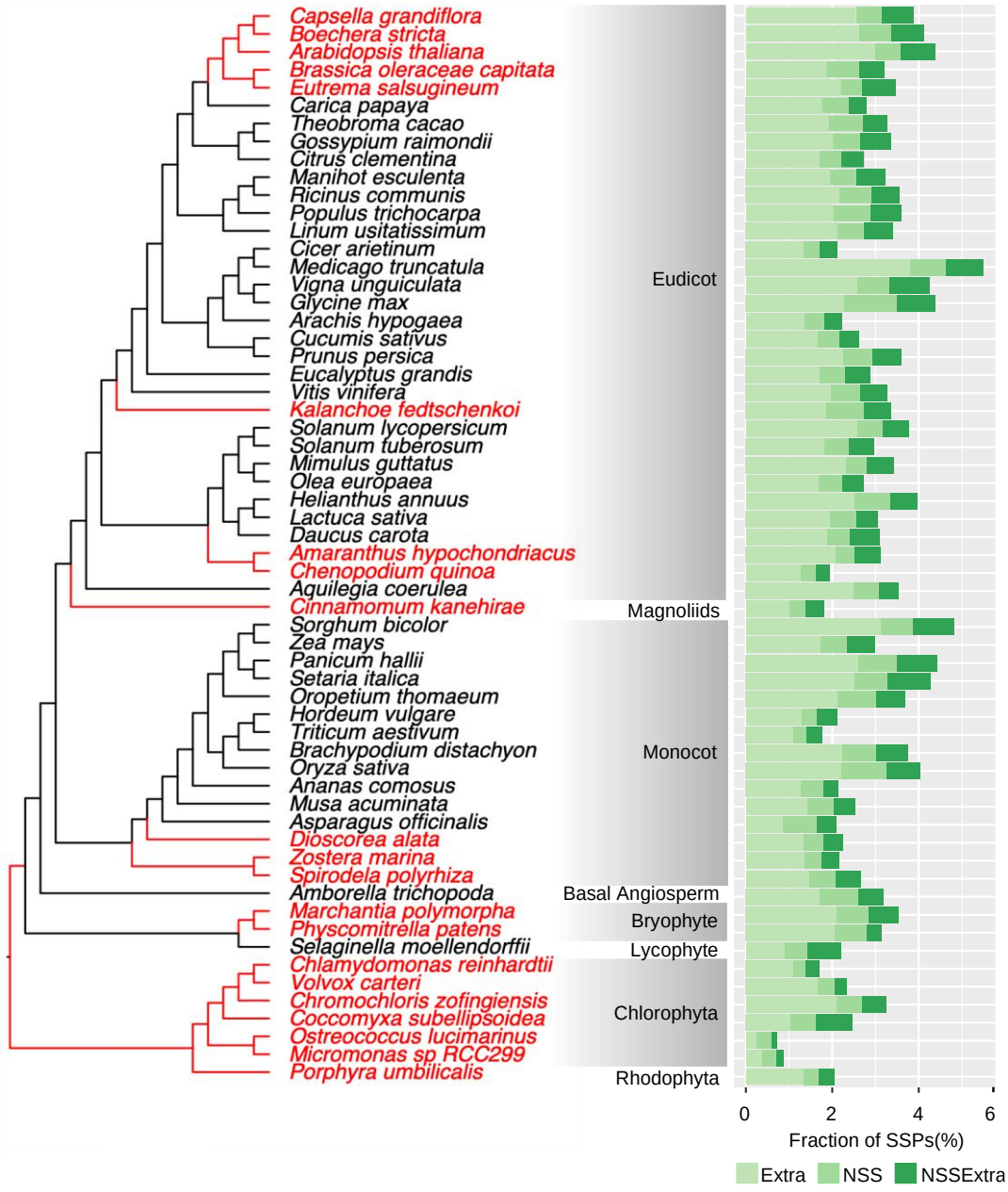

Figure 2.2 A coalescent-based maximum likelihood phylogenetic tree of 60 plant species inferred from single copy gene trees

Plant species with and without ability to form AMS are indicated in black and red, respectively. The bar plot on the right side of the phylogenetic tree indicates the fraction of predicted SSPs in each plant species. Extra represents SSPs in the “Extracellular-only set”; NSS represents SSPs in the “NSS-only set”; and NSSExtra represents SSPs in the “NSS-and-extracellular set”, as defined in Fig. 2.1.

encoding heavy-metal associated domain proteins), OG0009886 (containing genes encoding wall-associated receptor kinase), and OG0010641 (containing genes encoding protein with unknown function). Then from the ortholog groups containing SSPs from both AMS and non-AMS species, we identified three AMS-preferential ortholog groups (APOGs), in which the number of SSPs from the AMS species was significantly ($P < 0.05$) higher than that from the non-AMS species, including OG0000049 (containing genes encoding plastocyanin-like proteins), OG0000081 (containing genes encoding Dirigent proteins), and OG0000364 (containing genes encoding EPFL proteins). These AMS-specific and AMS-preferential genes were not found in the ancient plant lineages such as *Chromochloris zofingiensis*, *Chlamydomonas reinhardtii*, *Porphyra umbilicalis*, etc., and there were repeated emergence or expansion in multiple plant lineages (Fig. 2.3), suggesting that these AMS-associated genes resulted from convergent evolution.

AMF-regulated gene expression

To identify AMF-inducible SSPs, we performed a cross-species comparative analysis of gene expression in four AMS plant species (*C. sativus*, *M. esculenta*, *M. truncatula* and *T. aestivum*) and one non-AMS plant species (*A. thaliana*), which were inoculated with AMF *R. irregularis* (Table S2). Through analysis of differential gene expression between AMF treatments and corresponding controls at different time points after fungal inoculation, we identified a total of 45, 3,255, 8,582, 1,263, and 8,205 differentially expressed genes (DEGs) in *A. thaliana*, *C. sativus*, *M. esculenta*, *M. truncatula*, and *T. aestivum*, respectively (Table S4). To further explore if the expression of SSPs were affected by AMF, we checked the DEGs encoding SSPs in these species. We identified 91, 330, 47, and 193 differentially expressed SSPs in *C. sativus*, *M. esculenta*, *M. truncatula*, and *T. aestivum*, respectively. No differentially expressed SSPs were found in non-AMS *A. thaliana* (Table S5). Furthermore, we identified 27 and 34 ortholog groups containing SSPs that were up- and down-regulated, respectively, by AMF treatment in at least two of the four AMS species (Fig. 2.4), suggesting convergency in AMF-responsive gene expression among different plant species.

Diversification and conservation between genes in the AMS-preferential ortholog groups containing AMF-inducible SSPs

Three AMS-preferential ortholog groups (i.e., OG0000049, OG0000081, OG0000364) contained SSP genes up-regulated by AMF in at least two of the four AMS species. Divergence of protein functions are mainly determined by variations in 3D structure (Liu et al. 2019a). We performed 3D protein structural prediction for AMF-inducible SSPs and their closely related non-SSPs in the phylogenetic trees (Figs. S2, S3 and S4) of AMS-preferential ortholog groups containing AMF-inducible SSPs and found that the 3D structures of the SPPs and their related non-SSPs highly matched each other, with only

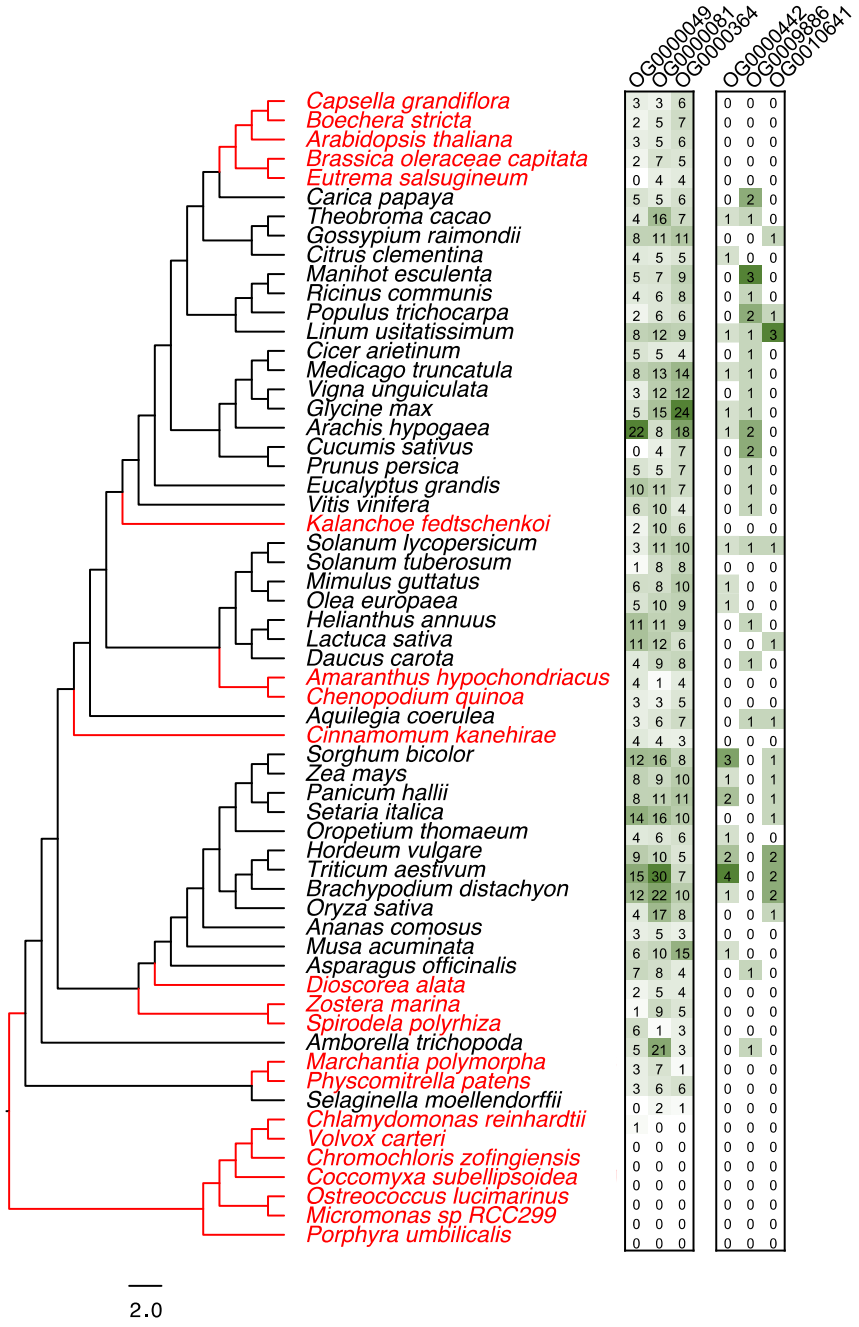

Figure 2.3 Number of small secreted proteins (SSPs) in representative ortholog groups.

OG0000049, OG0000081, and OG0000364 are AMS-preferential ortholog groups containing SSPs from at least 30% of the 39 AMS species. OG0000442, OG0009886, and OG0010641 are AMS-specific ortholog groups containing significantly ($P < 0.05$) more SSPs from the AMS species than from the non-AMS species. Relative abundance of SSPs within each ortholog group is represented by a color scale.

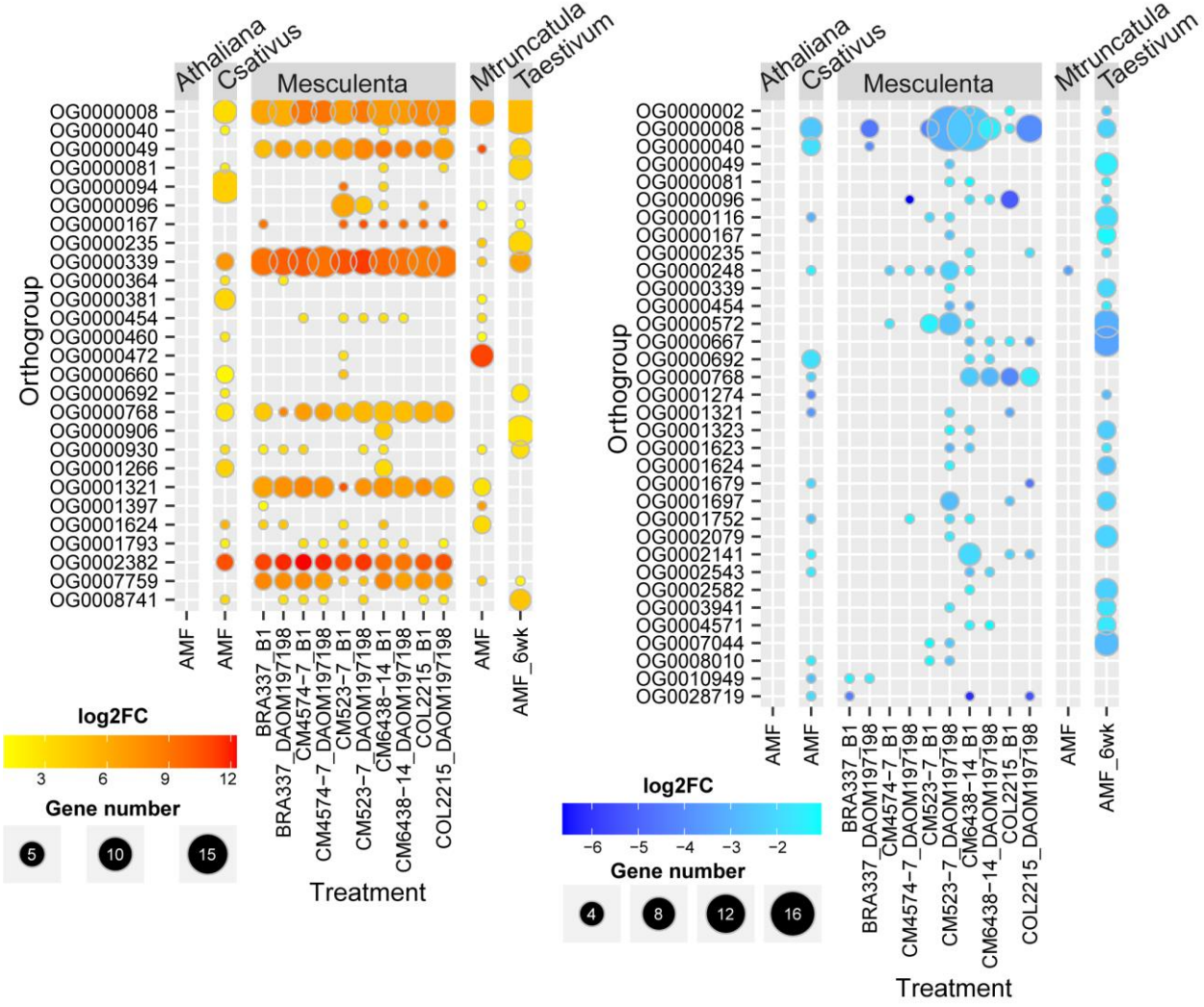

Figure 2.4 Ortholog groups containing small secreted proteins (SSPs) showing differential gene expression in response to AMF *Rhizophagus irregularis* in at least two plant species.

Upregulation and downregulation of plant SSP gene expression by the AMF treatment. The heatmap represents Log2 ratio of transcript abundance between AMF treatment versus control and the circle size indicates the number of SSPs in each ortholog group. The differential gene expression between AMF treatment and control was defined as at least two-fold change in transcript abundance, along with adjusted $p < 0.05$.

a few local variations (Fig. 2.5), suggesting that the evolution SSPs involves some minor structural changes.

Conserved *cis*-acting elements located in the gene promoter region regulate gene expression pattern (Liu et al. 2019a). We conducted comparative analysis of promoter sequences (i.e., 2000 bp upstream of the translation start codon) between various gene pairs selected from two AMS-preferential ortholog groups. Three *cis*-acting elements including the binding sites of transcription factors bHLH, GATA and MYB were found to be conserved in the promoter regions of SSP genes upregulated by AMF (Fig. 2.6). It has been reported that these transcription factors (bHLH, GATA and MYB) were involved in response to abiotic stresses, cell wall modification, and pathogens, respectively (Shikata et al. 2004; Lei et al. 2019; Jiang et al. 2020).

Co-expression analysis

To uncover additional context for potential function and evolutionary divergence of SSPs, the SSP co-expression networks were constructed by using woody model plant *Populus* as an example because it currently has a large amount of public gene expression datasets. To obtain the high confidential co-expression relationships, we extracted the highly co-expressed genes ($|PCC| \geq 0.95$) based on the *Populus* gene atlas. Finally, from 1248 SSPs in *Populus*, 353 SSPs were highly co-expressed with 34,980 genes. Then, we focused on the subnetworks of SSPs in AMS-specific ortholog groups (i.e., OG0000442, OG0009886, OG0010641) and AMS-preferential ortholog groups (i.e., OG0000049, OG0000081, OG0000364). Three genes in AMS-preferential ortholog groups OG0009886 and OG0010641 were co-expressed with 142 genes. Four genes (Potri.008G061400, Potri.016G060900, Potri.018G130700 and Potri.007G095400) in OG0000081 and OG0000364 were co-expressed with 99, 3, 2 and 1 genes, respectively (Fig. 2.7). The gene set co-expressed with Potri.008G061400, which encodes a disease resistance-responsive/dirigent-like protein, was overrepresented by genes involved in signalling, cell wall and stress (8, 5 and 4 genes), suggesting that Potri.008G061400 plays a role in diverse biological processes.

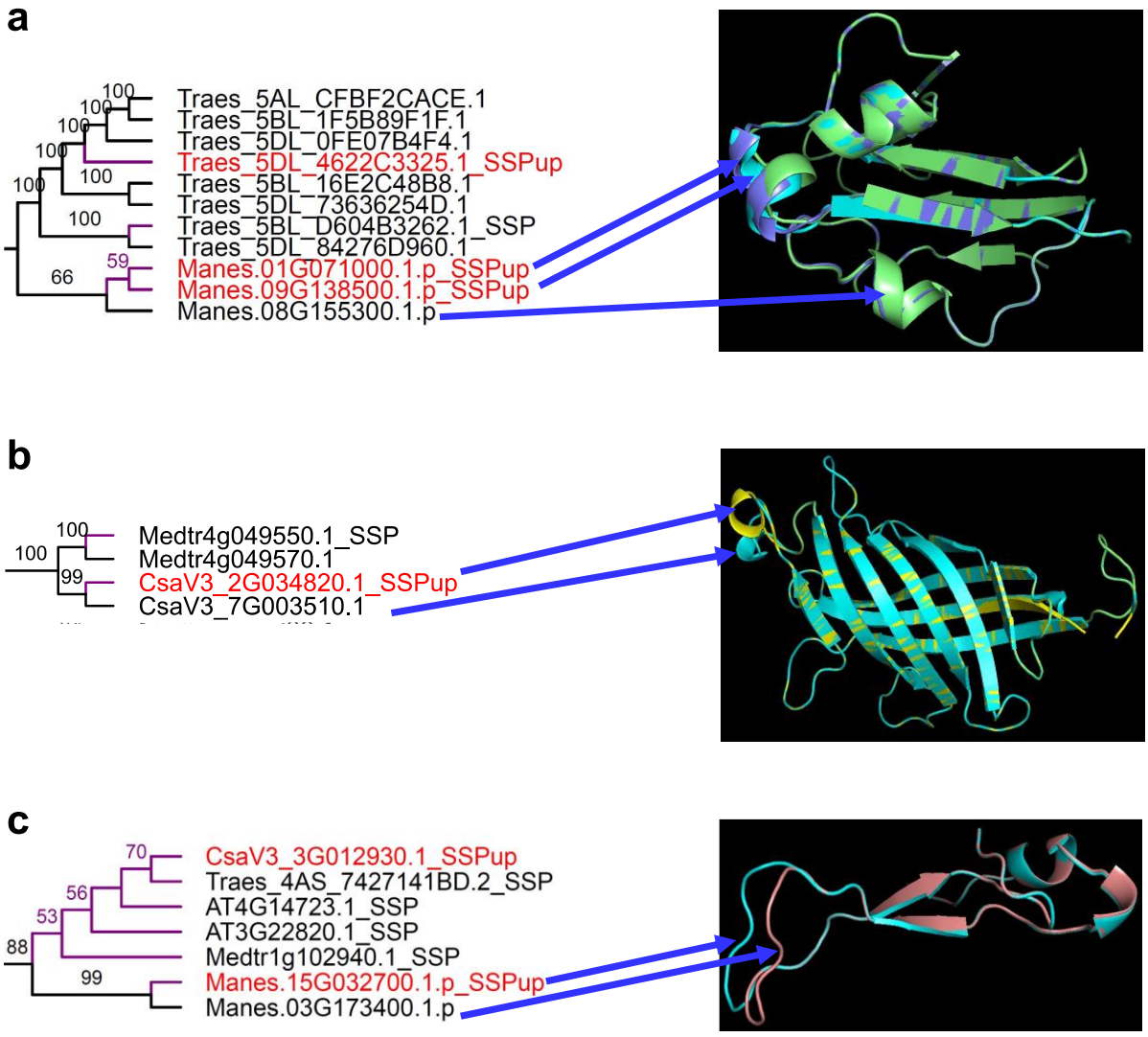

Figure 2.5 Structure modelling of AMS-related small secreted proteins (SSPs) and their closely related non-SSP sequences in the AMS-preferential ortholog groups.

Different colors indicate different proteins. Red arrows point out local variations found in protein structures in the AMS-preferential ortholog groups OG0000049 (**a**), OG0000081 (**b**), and OG0000364 (**c**).

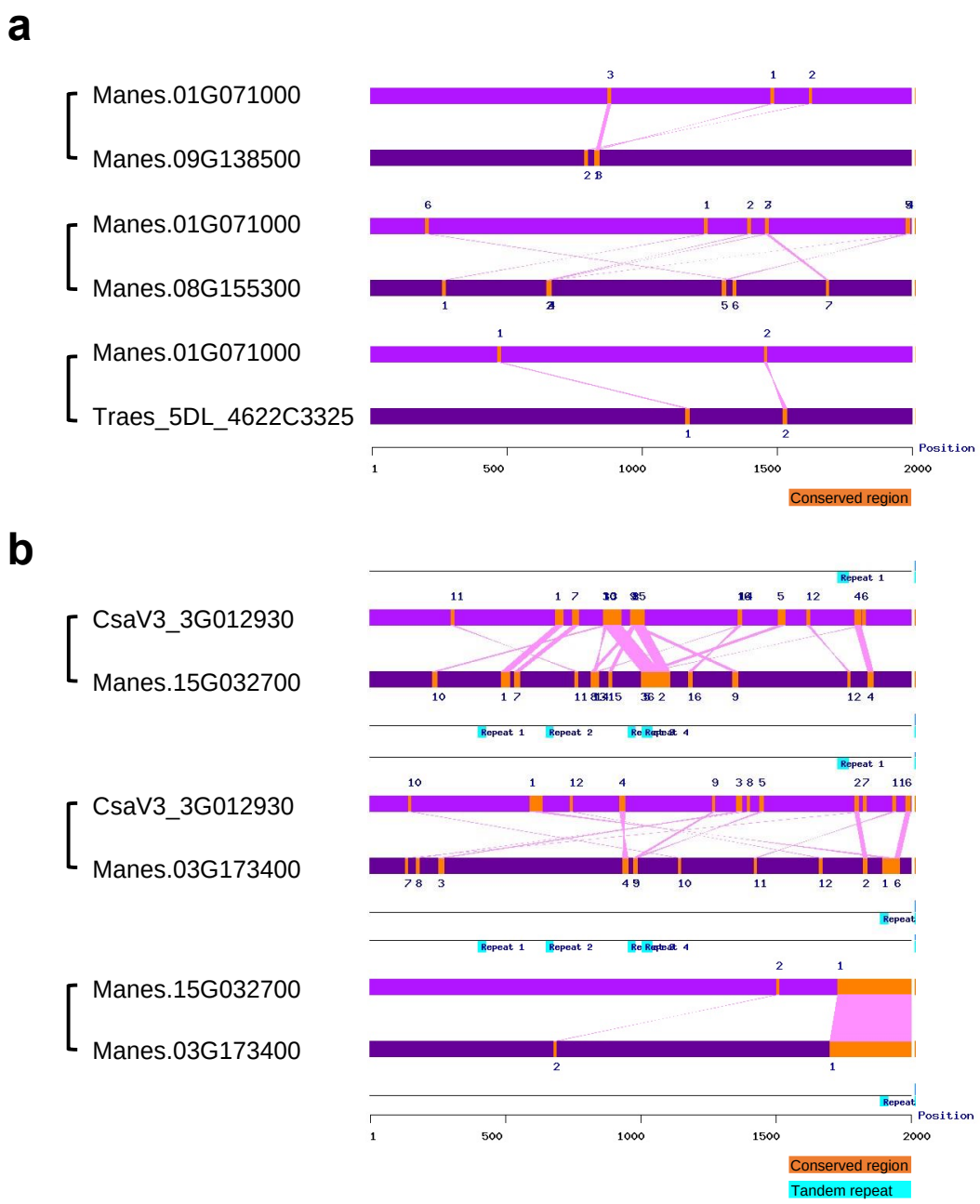

Figure 2.6 Promoter alignment between different gene pairs selected from AMS-preferential ortholog groups.

Conserved blocks were located in the promoter regions (i.e., 2000 bp upstream of the translation start codon) of AMF-inducible small secreted protein (SSP) genes, in comparison with closely related non-SSP genes, which are selected from AMS-preferential ortholog groups OG0000049 (**a**) and OG0000364 (**b**).

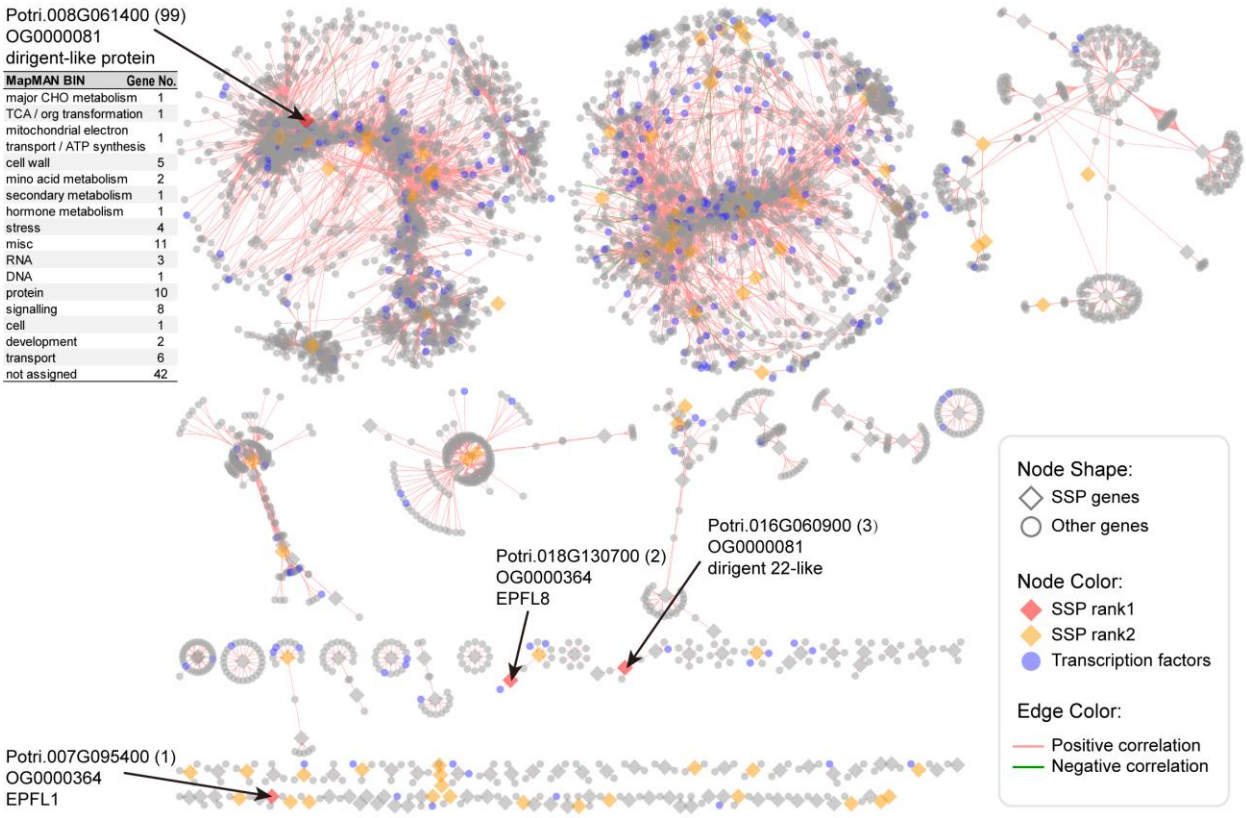

Figure 2.7 Co-expression network of *Populus trichocarpa* small secreted proteins (SSPs) in AMS-specific ortholog groups, AMS-preferential ortholog groups, and ortholog groups containing differential expressed SSPs from at least three species.

"SSP rank1" represents SSPs shared by the AMS-preferential ortholog groups and the orthogroups containing differential expressed SSPs from at least three species in response to AMF *Rhizophagus irregularis*. "SSP rank2" represents SSPs from the AMS-specific ortholog groups or the AMS-preferential ortholog groups or the ortholog groups containing differential expressed SSPs from at least three species in response to AMF.

Discussion

With increasing number of sequenced plant genomes and advancement in bioinformatics, more and more SSPs have been identified in various plants. However, there are several limitations in previous studies on SSP prediction. First, much attention has been paid on predicting NSS-containing SSPs, overlooking SSPs associated with unconventional secretion pathways. Second, most of previous efforts have relied upon single computational methods for predicting protein secretion, resulting in biased results because there is a big difference in the prediction result among different computational tools for protein secretion prediction (Figs. S1a, S1b and S1c). To reduce the false positive prediction of SSPs, we created a stringent workflow (Fig. 2.1) to predict SSPs, based on the consensus prediction of at least two of the three popular methods for predicting protein signal peptides or extra cellular localization.

Through comparative genomics analysis, we predicted AMS-related SSPs in AMS-specific ortholog groups (i.e., OG0000442, OG0009886, OG0010641) and AMS-preferential ortholog groups (i.e., OG0000049, OG0000081, OG0000364). The SSP genes in ortholog group OG0000049 encode glycosylphosphatidylinositol-anchored proteins (GPI-APs). GPI-APs are ubiquitous and abundant among eukaryotes (Kinoshita and Fujita 2016). To date, more than 300 GPI-APs have been identified in *A. thaliana*. These proteins are involved in signaling transduction during multiple biological processes, such as cell wall composition, hormone signaling responses and pathogen responses (Zhou 2019). In this study, we found that several SSP genes encoding disease resistance-responsive proteins in ortholog group OG0000081 were upregulated by AMF, suggesting that these SSPs could play roles in plant response to both pathogens and beneficial microbes.

Poplar (*Populus* spp.) is an important woody crop for bioenergy, horticulture, and ecosystems service (Dharmawardhana et al. 2009; Yang et al. 2009). Based on co-expression networks, we identified four *P. trichocarpa* genes (i.e., Potri.008G061400, Potri.016G060900, Potri.018G130700 and Potri.007G095400) in two AMS-preferential ortholog groups, which were co-expressed with other polar genes. For example, Potri.008G061400 encoding a disease resistance-responsive/dirigent-like protein is co-expressed with 99 polar genes with diverse functions (Fig. 2.7). This result suggests that SSPs can function in a complex network regulating multiple biological processes.

Convergent evolution plays an important role in plant-microbe interactions (Saijo et al. 2018; Carter et al. 2019; de Vries et al. 2020). Our phylogenomic analysis revealed that AMS emerged in multiple plant lineages through convergent evolution (Fig. 2.1). Through comparative genomics analysis, we found that some SSPs in the AMS-preferential ortholog groups showed convergent changes in gene expression in response to AMF (Fig. 2.4). Also, we found convergent emergency of SSPs in both the AMS-specific ortholog groups and the AMS-preferential ortholog groups (Fig. 2.3).

These results suggest that the convergent emergency of SSPs may play an important role in the convergent evolution of AMS.

CONCLUSION

The collection of work presented here includes a comprehensive summary regarding the current knowledge of plant SSPs and the attempt to explore the relation between plant SSPs and AMS. These studies provide a good background for other scientists to systematically understand plant SSPs and provide insight into evolutionary relationships between SSPs and AMS. The computational pipeline developed in this research can be applied for discovering SSPs in other plants species. The AMS-related SSP genes predicted in this work could serve as high-value candidates for experimental characterization to gain a deep understanding of the molecular mechanisms underlying the beneficial interactions between plants and AMF.

REFERENCES

Alexandersson E, Ashfaq A, Resjö S, Andreasson E. 2013. Plant secretome proteomics. *Frontiers in Plant Science* **4**: 9.

Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**: 3387-3395.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* **37**: 420-423.

Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics* **15**: 193-204.

Araya T, Miyamoto M, Wibowo J, Suzuki A, Kojima S, Tsuchiya YN, Sawa S, Fukuda H, Von Wirén N, Takahashi H. 2014. CLE-CLAVATA1 peptide-receptor signaling module regulates the expansion of plant root systems in a nitrogen-dependent manner. *Proceedings of the National Academy of Sciences* **111**: 2029-2034.

Armenteros JJA, Salvatore M, Emanuelsson O, Winther O, Von Heijne G, Elofsson A, Nielsen H. 2019a. Detecting sequence signals in targeting peptides using deep learning. *Life science alliance* **2**.

Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019b. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology* **37**: 420-423.

Atkinson NJ, Lilley CJ, Urwin PE. 2013. Identification of genes involved in the response of *Arabidopsis* to simultaneous biotic and abiotic stresses. *Plant Physiology* **162**: 2028-2041.

Bellion M, Courbot M, Jacob C, Blaudez D, Chalot M. 2006. Extracellular and cellular mechanisms sustaining metal tolerance in ectomycorrhizal fungi. *FEMS microbiology letters* **254**: 173-181.

Ben-Amar A, Daldoul S, M Reustle G, Krczal G, Mliki A. 2016. Reverse genetics and high throughput sequencing methodologies for plant functional genomics. *Current Genomics* **17**: 460-475.

Bobay BG, DiGennaro P, Scholl E, Imin N, Djordjevic MA, Mck Bird D. 2013. Solution NMR studies of the plant peptide hormone CEP inform function. *FEBS letters* **587**: 3979-3985.

Bolan N. 1991. A critical review on the role of mycorrhizal fungi in the uptake of phosphorus by plants. *Plant and soil* **134**: 189-207.

Bona E, Cantamessa S, Massa N, Manassero P, Marsano F, Copetta A, Lingua G, D'Agostino G, Gamalero E, Berta G. 2017. Arbuscular mycorrhizal fungi and plant growth-promoting pseudomonads improve yield, quality and nutritional value of tomato: a field study. *Mycorrhiza* **27**: 1-11.

Bonfante P, Genre A. 2010. Mechanisms underlying beneficial plant–fungus interactions in mycorrhizal symbiosis. *Nature communications* **1**: 1-11.

Boschiero C, Dai X, Lundquist PK, Roy S, Christian de Bang T, Zhang S, Zhuang Z, Torres-Jerez I, Udvardi MK, Scheible W-R et al. 2020. MtSSPdb: The *Medicago truncatula* small secreted peptide database. *Plant Physiology* **183**: 399-413.
Boschiero C, Lundquist PK, Roy S, Dai X, Zhao PX, Scheible W-R. 2019. Identification and functional investigation of genome-encoded, small, secreted peptides in plants. *Current Protocols in Plant Biology* **4**: e20098.
Brand U, Fletcher JC, Hobe M, Meyerowitz EM, Simon R. 2000. Dependence of stem cell fate in *Arabidopsis* on a feedback loop regulated by CLV3 activity. *Science* **289**: 617-619.
Briceño Z, Almagro L, Sabater-Jara AB, Calderón AA, Pedreño MA, Ferrer MA. 2012. Enhancement of phytosterols, taraxasterol and induction of extracellular pathogenesis-related proteins in cell cultures of *Solanum lycopersicum* cv Micro-Tom elicited with cyclodextrins and methyl jasmonate. *Journal of Plant Physiology* **169**: 1050-1058.
Briesemeister S, Rahnenführer J, Kohlbacher O. 2010. YLoc—an interpretable web server for predicting subcellular localization. *Nucleic Acids Research* **38**: W497-W502.
Brundrett MC. 2009. Mycorrhizal associations and other means of nutrition of vascular plants: understanding the global diversity of host plants by resolving conflicting information and developing reliable means of diagnosis. *Plant and Soil* **320**: 37-77.
Burko Y, Gaillochet C, Seluzicki A, Chory J, Busch W. 2020. Local HY5 activity mediates hypocotyl growth and shoot-to-root communication. *Plant Communications* **1**: 100078.
Cao B, Wei X-C, Xu X-R, Zhang H-Z, Luo C-H, Feng B, Xu R-C, Zhao S-Y, Du X-J, Han L. 2019. Seeing the unseen of the combination of two natural resins, frankincense and myrrh: Changes in chemical constituents and pharmacological activities. *Molecules* **24**: 3076.
Carter ME, Helm M, Chapman AV, Wan E, Restrepo Sierra AM, Innes RW, Bogdanove AJ, Wise RP. 2019. Convergent evolution of effector protease recognition by Arabidopsis and barley. *Molecular Plant-Microbe Interactions* **32**: 550-565.
Chae K, Lord EM. 2011. Pollen tube growth and guidance: roles of small, secreted proteins. *Annals of Botany* **108**: 627-636.
Chagas FO, Pessotti RC, Caraballo-Rodriguez AM, Pupo MT. 2018. Chemical signaling involved in plant-microbe interactions. *Chemical Society Reviews* **47**: 1652-1704.
Chan K-L, Rosli R, Tatarinova TV, Hogan M, Firdaus-Raih M, Low E-TL. 2017. Seqping: gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data. *BMC Bioinformatics* **18**: 1-7.
Cheli F, Baldi A. 2011. Nutrition-based health: Cell-based bioassays for food antioxidant activity evaluation. *Journal of Food Science* **76**: R197-R205.
Chen L, Zhai L, Li Y, Li N, Zhang C, Ping L, Chang L, Wu J, Li X, Shi D et al. 2015. Development of gel-filter method for high enrichment of low-molecular weight proteins from serum. *PloS One* **10**: e0115862-e0115862.

Chen X, Yao Q, Gao X, Jiang C, Harberd NP, Fu X. 2016. Shoot-to-root mobile transcription factor HY5 coordinates plant carbon and nitrogen acquisition. *Current Biology* **26**: 640-646.
Chen Y-L, Lee C-Y, Cheng K-T, Chang W-H, Huang R-N, Nam HG, Chen Y-R. 2014. Quantitative peptidomics study reveals that a wound-induced peptide from PR-1 regulates immune signaling in tomato. *The Plant Cell* **26**: 4135-4148.
Chen YL, Fan KT, Hung SC, Chen YR. 2020. The role of peptides cleaved from protein precursors in eliciting plant stress reactions. *New Phytologist* **225**: 2267-2282.
Cheng Q, Cao Y, Jiang C, Xu La, Wang M, Zhang S, Huang M. 2010. Identifying secreted proteins of *Marssonina brunnea* by degenerate PCR. *Proteomics* **10**: 2406-2417.
Chow C-N, Lee T-Y, Hung Y-C, Li G-Z, Tseng K-C, Liu Y-H, Kuo P-L, Zheng H-Q, Chang W-C. 2019. PlantPAN3. 0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic acids research* **47**: D1155-D1163.
Chuang C-F, Meyerowitz EM. 2000. Specific and heritable genetic interference by double-stranded RNA in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences* **97**: 4985-4990.
Clark SE, Running MP, Meyerowitz EM. 1995. CLAVATA3 is a specific regulator of shoot and floral meristem development affecting the same processes as CLAVATA1. *Development* **121**: 2057-2067.
Constabel CP, Yip L, Ryan CA. 1998. Prosystemin from potato, black nightshade, and bell pepper: primary structure and biological activity of predicted systemin polypeptides. *Plant Molecular Biology* **36**: 55-62.
Covey PA, Subbaiah CC, Parsons RL, Pearce G, Lay FT, Anderson MA, Ryan CA, Bedinger PA. 2010. A pollen-specific RALF from tomato that regulates pollen tube elongation. *Plant Physiology* **153**: 703-715.
de Bang TC, Lundquist PK, Dai X, Boschiero C, Zhuang Z, Pant P, Torres-Jerez I, Roy S, Nogales J, Veerappan V. 2017. Genome-wide identification of *Medicago* peptides involved in macronutrient responses and nodulation. *Plant Physiology* **175**: 1669-1689.
De Smet I, Vassileva V, De Rybel B, Levesque MP, Grunewald W, Van Damme D, Van Noorden G, Naudts M, Van Isterdael G, De Clercq R. 2008. Receptor-like kinase ACR4 restricts formative cell divisions in the *Arabidopsis* root. *Science* **322**: 594-597.
de Vries S, Stukenbrock EH, Rose LE. 2020. Rapid evolution in plant–microbe interactions–an evolutionary genomics perspective. *New Phytologist* **226**: 1256-1262.
De-la-Pena C, Badri DV, Loyola-Vargas VM. 2012. Plant root secretions and their interactions with neighbors. In *Secretions and exudates in biological systems*, pp. 1-26. Springer.
De-la-Peña C, M Loyola-Vargas V. 2012. The hidden chemical cross-talk between roots and microbes: a proteomic approach. *Current Proteomics* **9**: 103-117.

Delaux PM. 2017. Comparative phylogenomics of symbiotic associations. *New Phytologist* **213**: 89-94.
Demarque DP, Dusi RG, de Sousa FD, Grossi SM, Silvério MR, Lopes NP, Espindola LS. 2020. Mass spectrometry-based metabolomics approach in the isolation of bioactive natural products. *Scientific Reports* **10**: 1-9.
Dharmawardhana P, Brunner A, Strauss S. 2009. Poplar as a Tree Model for Horticulture and Beyond: a Case Study of Genome-Scale Changes in Gene Expression during Bud Entry and Release from Dormancy. In *International Symposium on Molecular Markers in Horticulture 859*, pp. 43-47.
Ding M, Tegel H, Sivertsson Å, Hober S, Snijder A, Ormö M, Strömstedt P-E, Davies R, Holmberg Schiavone L. 2020. Secretome-based screening in target discovery. *Slas Discovery* **25**: 535-551.
Ding Y, Robinson DG, Jiang L. 2014. Unconventional protein secretion (UPS) pathways in plants. *Current Opinion in Cell Biology* **29**: 107-115.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.
Elorriaga E, Klocko AL, Ma C, Strauss SH. 2018. Variation in mutation spectra among CRISPR/Cas9 mutagenized poplars. *Frontiers in Plant Science* **9**: 594.
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology* **20**: 1-14.
Erffelinck M-L, Ribeiro B, Perassolo M, Pauwels L, Pollier J, Storme V, Goossens A. 2018. A user-friendly platform for yeast two-hybrid library screening using next generation sequencing. *PloS one* **13**: e0201270.
Etchells JP, Turner SR. 2010. The PXY-CLE41 receptor ligand pair defines a multifunctional pathway that controls the rate and orientation of vascular cell division. *Development* **137**: 767-774.
Farkas A, Maróti G, Dürgő H, Györgypál Z, Lima RM, Medzihradszky KF, Kereszt A, Mergaert P, Kondorosi É. 2014. Medicago truncatula symbiotic peptide NCR247 contributes to bacteroid differentiation through multiple mechanisms. *Proceedings of the National Academy of Sciences* **111**: 5183-5188.
Fernandez A, Drozdzecki A, Hoogewijs K, Vassileva V, Madder A, Beeckman T, Hilson P. 2015. The GLV6/RGF8/CLEL2 peptide regulates early pericycle divisions during lateral root initiation. *Journal of Experimental Botany* **66**: 5245-5256.
Fletcher JC, Brand U, Running MP, Simon R, Meyerowitz EM. 1999. Signaling of cell fate decisions by CLAVATA3 in *Arabidopsis* shoot meristems. *Science* **283**: 1911-1914.
Fukuda H, Hardtke CS. 2020. Peptide signaling pathways in vascular differentiation. *Plant Physiology* **182**: 1636.
Fukuda H, Ohashi-Ito K. 2019. Vascular tissue development in plants. *Current Topics in Developmental Biology* **131**: 141-160.
Ghorbani S, Lin Y-C, Parizot B, Fernandez A, Njo MF, Van de Peer Y, Beeckman T, Hilson P. 2015. Expanding the repertoire of secretory peptides controlling root

development with comparative genome analysis and functional assays. *Journal of Experimental Botany* **66**: 5257-5269.
Gilchrist E, Haughn G. 2010. Reverse genetics techniques: engineering loss and gain of gene function in plants. *Briefings in Functional Genomics* **9**: 103-110.
Goldberg T, Hamp T, Rost B. 2012. LocTree2 predicts localization for all domains of life. *Bioinformatics* **28**: i458-i465.
Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, Altermann U, Angerer P, Ansorge S, Balasz K et al. 2014. LocTree3 prediction of localization. *Nucleic Acids Research* **42**: W350-W355.
Gonzalez-Rizzo S, Crespi M, Frugier F. 2006. The *Medicago truncatula* CRE1 cytokinin receptor regulates lateral root development and early symbiotic interaction with *Sinorhizobium meliloti*. *The Plant Cell* **18**: 2680-2693.
Goring DR, Di Sansebastiano GP. 2018. Protein and membrane trafficking routes in plants: conventional or unconventional? *Journal of Experimental Botany* **69**: 1–5.
Greening DW, Simpson RJ. 2010. A centrifugal ultrafiltration strategy for isolating the low-molecular weight (⩽25K) component of human plasma proteome. *Journal of Proteomics* **73**: 637-648.
Guo J-C, Fang S-S, Wu Y, Zhang J-H, Chen Y, Liu J, Wu B, Wu J-R, Li E-M, Xu L-Y. 2019. CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic acids research* **47**: W516-W522.
Gupta R, Deswal R. 2012. Low temperature stress modulated secretome analysis and purification of antifreeze protein from *Hippophae rhamnoides*, a Himalayan wonder plant. *Journal of Proteome Research* **11**: 2684-2696.
Han S, Khan MHU, Yang Y, Zhu K, Li H, Zhu M, Amoo O, Khan S, Fan C, Zhou Y. 2020. Identification and comprehensive analysis of the CLV3/ESR-related (CLE) gene family in *Brassica napus* L. *Plant Biology* **22**: 709-721.
Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, Shiu S-H. 2010. sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26**: 399-400.
Handa Y, Nishide H, Takeda N, Suzuki Y, Kawaguchi M, Saito K. 2015. RNA-seq transcriptional profiling of an arbuscular mycorrhiza provides insights into regulated and coordinated gene expression in *Lotus japonicus* and *Rhizophagus irregularis*. *Plant and Cell Physiology* **56**: 1490-1511.
Hara K, Kajita R, Torii KU, Bergmann DC, Kakimoto T. 2007. The secretory peptide gene EPF1 enforces the stomatal one-cell-spacing rule. *Genes & development* **21**: 1720-1725.
Hassan MM, Yuan G, Chen J-G, Tuskan GA, Yang X. 2020. Prime editing technology and its prospects for future applications in plant biology research. *BioDesign Research* **2020**: 9350905.
Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC. 2016. The emerging world of small ORFs. *Trends in Plant Science* **21**: 317-328.
Horváth B, Domonkos Á, Kereszt A, Szűcs A, Ábrahám E, Ayaydin F, Bóka K, Chen Y, Chen R, Murray JD. 2015. Loss of the nodule-specific cysteine rich peptide,

NCR169, abolishes symbiotic nitrogen fixation in the Medicago truncatula dnf7 mutant. *Proceedings of the National Academy of Sciences* **112**: 15232-15237.
Hou S, Wang X, Chen D, Yang X, Wang M, Turrà D, Di Pietro A, Zhang W. 2014. The secreted peptide PIP1 amplifies immunity through receptor-like kinase 7. *PLoS Pathog* **10**: e1004331.
Hsu PY, Benfey PN. 2018. Small but mighty: functional peptides encoded by small ORFs in plants. *Proteomics* **18**: 1700038.
Hu XL, Lu H, Hassan MM, Zhang J, Yuan G, Abraham PE, Shrestha HK, Villalobos Solis MI, Chen JG, Tschaplinski TJ et al. 2021. Advances and perspectives in discovery and functional analysis of small secreted proteins in plants. *Hortic Res* **8**: 130.
Huffaker A, Pearce G, Ryan CA. 2006. An endogenous peptide signal in Arabidopsis activates components of the innate immune response. *Proceedings of the National Academy of Sciences* **103**: 10098-10103.
Hunt L, Bailey KJ, Gray JE. 2010. The signalling peptide EPFL9 is a positive regulator of stomatal development. *New Phytologist*: 609-614.
Imin N, Mohd-Radzman NA, Ogilvie HA, Djordjevic MA. 2013. The peptide-encoding CEP1 gene modulates lateral root and nodule numbers in *Medicago truncatula*. *Journal of Experimental Botany* **64**: 5395-5409.
Ito Y, Nakanomyo I, Motose H, Iwamoto K, Sawa S, Dohmae N, Fukuda H. 2006. Dodeca-CLE peptides as suppressors of plant stem cell differentiation. *Science* **313**: 842-845.
Jiang B, Shi Y, Peng Y, Jia Y, Yan Y, Dong X, Li H, Dong J, Li J, Gong Z. 2020. Cold-induced CBF–PIF3 interaction enhances freezing tolerance by stabilizing the phyB thermosensor in Arabidopsis. *Molecular plant* **13**: 894-906.
Johnson NC, Graham J, Smith F. 1997. Functioning of mycorrhizal associations along the mutualism–parasitism continuum. *The New Phytologist* **135**: 575-585.
Jones DL, Nguyen C, Finlay RD. 2009. Carbon flow in the rhizosphere: carbon trading at the soil–root interface. *Plant and soil* **321**: 5-33.
Käll L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic acids research* **35**: W429-W432.
Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**: 587-589.
Kandoth PK, Ranf S, Pancholi SS, Jayanty S, Walla MD, Miller W, Howe GA, Lincoln DE, Stratmann JW. 2007. Tomato MAPKs LeMPK1, LeMPK2, and LeMPK3 function in the systemin-mediated defense response against herbivorous insects. *Proceedings of the National Academy of Sciences* **104**: 12205-12210.
Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics* **20**: 1160-1166.
Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**: 845-858.

Kim Y-G, Lone AM, Saghatelian A. 2013. Analysis of the proteolysis of bioactive peptides using a peptidomics approach. *Nature Protocols* **8**: 1730.
Kinoshita A, Betsuyaku S, Osakabe Y, Mizuno S, Nagawa S, Stahl Y, Simon R, Yamaguchi-Shinozaki K, Fukuda H, Sawa S. 2010. RPK2 is an essential receptor-like kinase that transmits the CLV3 signal in *Arabidopsis*. *Development* **137**: 3911-3920.
Kinoshita T, Fujita M. 2016. Biosynthesis of GPI-anchored proteins: special emphasis on GPI lipid remodeling. *Journal of lipid research* **57**: 6-24.
Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canbäck B, Choi C, Cichocki N, Clum A et al. 2015. Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nature Genetics* **47**: 410-415.
Kondo T, Sawa S, Kinoshita A, Mizuno S, Kakimoto T, Fukuda H, Sakagami Y. 2006. A plant peptide encoded by CLV3 identified by in situ MALDI-TOF MS analysis. *Science* **313**: 845-848.
Krause C, Richter S, Knöll C, Jürgens G. 2013. Plant secretome - From cellular process to biological activity. *Biochimica et Biophysica Acta* **1834**: 2429-2441.
Kucukoglu M, Chaabouni S, Zheng B, Mähönen AP, Helariutta Y, Nilsson O. 2020. Peptide encoding *Populus CLV3/ESR-RELATED 47* (*PttCLE47*) promotes cambial development and secondary xylem formation in hybrid aspen. *New Phytologist* **226**: 75-85.
Laffont C, Ivanovici A, Gautrat P, Brault M, Djordjevic MA, Frugier F. 2020. The NIN transcription factor coordinates CEP and CLE signaling peptides that regulate nodulation antagonistically. *Nature Communications* **11**: 1-13.
Lanfranco L, Fiorilli V, Gutjahr C. 2018. Partner communication and role of nutrients in the arbuscular mycorrhizal symbiosis. *New Phytologist* **220**: 1031-1046.
Lauressergues D, Couzigou J-M, San Clemente H, Martinez Y, Dunand C, Bécard G, Combier J-P. 2015. Primary transcripts of microRNAs encode regulatory peptides. *Nature* **520**: 90-93.
Lease KA, Walker JC. 2006. The *Arabidopsis* unannotated secreted peptide database, a resource for plant peptidomics. *Plant Physiology* **142**: 831-838.
Lehto T. 1992. Mycorrhizas and drought resistance of *Picea sitchensis* (Bong.) Carr. I. In conditions of nutrient deficiency. *New phytologist* **122**: 661-668.
Lei J, Jayaprakasha GK, Singh J, Uckoo R, Borrego EJ, Finlayson S, Kolomiets M, Patil BS, Braam J, Zhu-Salzman K. 2019. CIRCADIAN CLOCK-ASSOCIATED1 controls resistance to aphids by altering indole glucosinolate production. *Plant physiology* **181**: 1344-1359.
Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C. 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**: 1035-1043.
Li J, Li Y, Ma L. 2019. CRISPR/Cas9-based genome editing and its applications for functional genomic analyses in plants. *Small Methods* **3**: 1800473.

Li YL, Dai XR, Yue X, Gao X-Q, Zhang XS. 2014. Identification of small secreted peptides (SSPs) in maize and expression analysis of partial SSP genes in reproductive tissues. *Planta* **240**: 713-728.
Liu B, Hu J, Zhang J. 2019a. Evolutionary divergence of duplicated Hsf genes in Populus. *Cells* **8**: 438.
Liu D, Chen M, Mendoza B, Cheng H, Hu R, Li L, Trinh CT, Tuskan GA, Yang X. 2019b. CRISPR/Cas9-mediated targeted mutagenesis for functional genomics research of crassulacean acid metabolism plants. *Journal of Experimental Botany* **70**: 6621-6629.
Liu D, Hu R, Palla KJ, Tuskan GA, Yang X. 2016. Advances and perspectives on the use of CRISPR/Cas9 systems in plant genomics research. *Current Opinion in Plant Biology* **30**: 70-77.
Liu D, Mewalal R, Hu R, Tuskan GA, Yang X. 2017. New technologies accelerate the exploration of non-coding RNAs in horticultural plants. *Horticulture Research* **4**: 17031.
Liu Y, Joly V, Dorion S, Rivoal J, Matton DP. 2015. The plant ovule secretome: A different view toward pollen-pistil interactions. *Journal of Proteome Research* **14**: 4763-4775.
Lowder LG, Paul JW, Qi Y. 2017. Multiplexed transcriptional activation or repression in plants using CRISPR-dCas9-based systems. In *Plant Gene Regulatory Networks Methods in Molecular Biology*, Vol 1629 (ed. K Kaufmann, B Mueller-Roeber), pp. 167-184. Humana Press, New York, NY.
Lum G, Meinken J, Orr J, Frazier S, Min XJ. 2014. PlantSecKB: the plant secretome and subcellular proteome knowledgebase. *Computational Molecular Biology* **4**: 1-17.
Ma B, Johnson R. 2012. *De novo* sequencing and homology searching. *Molecular & Cellular Proteomics* **11**: O111. 014902.
Mabona U, Viljoen A, Shikanga E, Marston A, Van Vuuren S. 2013. Antimicrobial activity of southern African medicinal plants with dermatological relevance: from an ethnopharmacological screening approach, to combination studies and the isolation of a bioactive compound. *Journal of Ethnopharmacology* **148**: 45-55.
MacLean AM, Bravo A, Harrison MJ. 2017. Plant signaling and metabolic pathways enabling arbuscular mycorrhizal symbiosis. *The Plant Cell* **29**: 2319-2335.
Makarewich CA, Olson EN. 2017. Mining for micropeptides. *Trends in Cell Biology* **27**: 685-696.
Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, Saghatelian A. 2020. Accurate annotation of human protein-coding small open reading frames. *Nature Chemical Biology* **16**: 458-468.
Matsubayashi Y, Sakagami Y. 1996. Phytosulfokine, sulfated peptides that induce the proliferation of single mesophyll cells of *Asparagus officinalis* L. *Proceedings of the National Academy of Sciences* **93**: 7623-7627.
Matsuzaki Y, Ogawa-Ohnishi M, Mori A, Matsubayashi Y. 2010. Secreted peptide signals required for maintenance of root stem cell niche in Arabidopsis. *Science* **329**: 1065-1067.

Meng L, Feldman LJ. 2010. CLE14/CLE20 peptides may interact with CLAVATA2/CORYNE receptor-like kinases to irreversibly inhibit cell division in the root meristem of *Arabidopsis*. *Planta* **232**: 1061-1074.
Mewalal R, Yin H, Hu R, Jawdy S, Vion P, Tuskan GA, Le Tacon F, Labbé JL, Yang X. 2019. Identification of Populus small RNAs responsive to mutualistic interactions with mycorrhizal fungi, Laccaria bicolor and Rhizophagus irregularis. *Frontiers in microbiology* **10**: 515.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution* **37**: 1530-1534.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**: i541-i548.
Mishima M, Takayama S, Sasaki K-i, Jee J-g, Kojima C, Isogai A, Shirakawa M. 2003. Structure of the male determinant factor for *Brassica* self-incompatibility. *Journal of Biological Chemistry* **278**: 36389-36395.
Mohd-Radzman NA, Binos S, Truong TT, Imin N, Mariani M, Djordjevic MA. 2015. Novel MtCEP1 peptides produced in vivo differentially regulate root development in *Medicago truncatula*. *Journal of Experimental Botany* **66**: 5289-5300.
Möller S, Croning MD, Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**: 646-653.
Moroder L, Musiol HJ, Götz M, Renner C. 2005. Synthesis of single- and multiple-stranded cystine-rich peptides. *Biopolymers* **80**: 85-97.
Mosher S, Seybold H, Rodriguez P, Stahl M, Davies KA, Dayaratne S, Morillo SA, Wierzba M, Favery B, Keller H. 2013. The tyrosine‐sulfated peptide receptors PSKR1 and PSY1R modify the immunity of Arabidopsis to biotrophic and necrotrophic pathogens in an antagonistic manner. *The Plant Journal* **73**: 469-482.
Müller R, Bleckmann A, Simon R. 2008. The receptor kinase CORYNE of *Arabidopsis* transmits the stem cell-limiting signal CLAVATA3 independently of CLAVATA1. *The Plant Cell* **20**: 934-946.
Murphy E, Smith S, De Smet I. 2012. Small signaling peptides in *Arabidopsis* development: how cells communicate over a short distance. *The Plant Cell* **24**: 3198-3217.
Nakaminami K, Okamoto M, Higuchi-Takeuchi M, Yoshizumi T, Yamaguchi Y, Fukao Y, Shimizu M, Ohashi C, Tanaka M, Matsui M. 2018. AtPep3 is a hormone-like peptide that plays a role in the salinity stress tolerance of plants. *Proceedings of the National Academy of Sciences* **115**: 5810-5815.
Ngcala MG, Goche T, Brown AP, Chivasa S, Ngara R. 2020. Heat stress triggers differential protein accumulation in the extracellular matrix of sorghum cell suspension cultures. *Proteomes* **8**: 29.
Nguyen TT, Lee H-H, Park J, Park I, Seo Y-S. 2017. Computational identification and comparative analysis of secreted and transmembrane proteins in six *Burkholderia* species. *Plant Pathol J* **33**: 148-162.

Nielsen H, Petsalaki EI, Zhao L, Stühler K. 2019. Predicting eukaryotic protein secretion without signals. *Biochimica et Biophysica Acta* **1867**: 140174.
Norkunas K, Harding R, Dale J, Dugdale B. 2018. Improving agroinfiltration-based transient gene expression in *Nicotiana benthamiana*. *Plant Methods* **14**: 71.
Nugent T, Jones DT. 2012. Detecting pore-lining regions in transmembrane protein sequences. *BMC Bioinformatics* **13**: 1-9.
Nwachukwu ID, Aluko RE. 2019. Structural and functional properties of food protein-derived antioxidant peptides. *Journal of Food Biochemistry* **43**: e12761.
Ohki S, Takeuchi M, Mori M. 2011. The NMR structure of stomagen reveals the basis of stomatal density regulation by plant peptide hormones. *Nature Communications* **2**: 1-7.
Ohyama K, Ogawa M, Matsubayashi Y. 2008a. Identification of a biologically active, small, secreted peptide in *Arabidopsis* by in silico gene screening, followed by LC-MS-based structure analysis. *The Plant Journal* **55**: 152-160.
Ohyama K, Ogawa M, Matsubayashi Y. 2008b. Identification of a biologically active, small, secreted peptide in *Arabidopsis* by in silico gene screening, followed by LC-MS-based structure analysis. *Plant J* **55**: 152-160.
Pan B, Sheng J, Sun W, Zhao Y, Hao P, Li X. 2012. OrysPSSP: a comparative Platform for Small Secreted Proteins from rice and other plants. *Nucleic Acids Research* **41**: D1192-D1198.
Patel N, Mohd-Radzman NA, Corcilius L, Crossett B, Connolly A, Cordwell SJ, Ivanovici A, Taylor K, Williams J, Binos S. 2018. Diverse peptide hormones affecting root growth identified in the *Medicago truncatula* secreted peptidome. *Molecular & Cellular Proteomics* **17**: 160-174.
Pearce G, Moura DS, Stratmann J, Ryan CA. 2001. Production of multiple plant hormones from a single polyprotein precursor. *Nature* **411**: 817-820.
Pearce G, Strydom D, Johnson S, Ryan CA. 1991. A polypeptide from tomato leaves induces wound-inducible proteinase inhibitor proteins. *Science* **253**: 895-897.
Peeters MK, Menschaert G. 2020. The hunt for sORFs: a multidisciplinary strategy. *Experimental cell research* **391**: 111923.
Péret B, Larrieu A, Bennett MJ. 2009. Lateral root emergence: a difficult birth. *Journal of Experimental Botany* **60**: 3637-3643.
Pinedo M, Regente M, Elizalde M, Y Quiroga I, A Pagnussat L, Jorrin-Novo J, Maldonado A, de la Canal L. 2012. Extracellular sunflower proteins: evidence on non-classical secretion of a jacalin-related lectin. *Protein and Peptide Letters* **19**: 270-276.
Plett JM, Daguerre Y, Wittulsky S, Vayssières A, Deveau A, Melton SJ, Kohler A, Morrell-Falvey JL, Brun A, Veneault-Fourrey C. 2014. Effector MiSSP7 of the mutualistic fungus Laccaria bicolor stabilizes the Populus JAZ6 protein and represses jasmonic acid (JA) responsive genes. *Proceedings of the National Academy of Sciences*: 201322671.
Plett JM, Yin H, Mewalal R, Hu R, Li T, Ranjan P, Jawdy S, De Paoli HC, Butler G, Burch-Smith TM et al. 2017. *Populus trichocarpa* encodes small, effector-like

secreted proteins that are highly induced during mutualistic symbiosis. *Scientific Reports* **7**: 382.
Potocka I, Baldwin TC, Kurczynska EU. 2012. Distribution of lipid transfer protein 1 (LTP1) epitopes associated with morphogenic events during somatic embryogenesis of Arabidopsis thaliana. *Plant cell reports* **31**: 2031-2045.
Rahmani F, Hummel M, Schuurmans J, Wiese-Klinkenberg A, Smeekens S, Hanson J. 2009. Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. *Plant Physiology* **150**: 1356-1367.
Rao VS, Srinivas K, Sujini G, Kumar G. 2014. Protein-protein interaction detection: methods and analysis. *International Journal of Proteomics* **2014**.
Read D, Perez‐Moreno J. 2003. Mycorrhizas and nutrient cycling in ecosystems‐a journey towards relevance? *New Phytologist* **157**: 475-492.
Rillig MC, Aguilar‐Trigueros CA, Camenzind T, Cavagnaro TR, Degrune F, Hohmann P, Lammel DR, Mansour I, Roy J, van der Heijden MG. 2019. Why farmers should manage the arbuscular mycorrhizal symbiosis. *New Phytologist* **222**: 1171-1175.
Rodriguez-Furlan C, Raikhel NV, Hicks GR. 2018. Merging roads: chemical tools and cell biology to study unconventional protein secretion. *Journal of Experimental Botany* **69**: 39-46.
Röhrig H, Schmidt J, Miklashevichs E, Schell J, John M. 2002. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proceedings of the National Academy of Sciences* **99**: 1915-1920.
Rojo E, Sharma VK, Kovaleva V, Raikhel NV, Fletcher JC. 2002. CLV3 is localized to the extracellular space, where it activates the *Arabidopsis* CLAVATA stem cell signaling pathway. *The Plant Cell* **14**: 969-977.
Ross A, Yamada K, Hiruma K, Yamashita‐Yamada M, Lu X, Takano Y, Tsuda K, Saijo Y. 2014. The Arabidopsis PEPR pathway couples local and systemic plant immunity. *The EMBO journal* **33**: 62-75.
Runyoro DK, Matee MI, Ngassapa OD, Joseph CC, Mbwambo ZH. 2006. Screening of Tanzanian medicinal plants for anti-Candida activity. *BMC Complementary and Alternative Medicine* **6**: 1-10.
Rutter BD, Innes RW. 2017. Extracellular vesicles isolated from the leaf apoplast carry stress-response proteins. *Plant Physiology* **173**: 728-741.
Sagaram US, El-Mounadi K, Buchko GW, Berg HR, Kaur J, Pandurangi RS, Smith TJ, Shah DM. 2013. Structural and functional studies of a phosphatidic acid-binding antifungal plant defensin MtDef4: identification of an RGFRRR motif governing fungal cell entry. *PLoS One* **8**: e82485.
Sahu SS, Loaiza CD, Kaundal R. 2020. Plant-mSubP: a computational framework for the prediction of single-and multi-target protein subcellular localization using integrated machine-learning approaches. *AoB Plants* **12**: plz068.
Saijo Y, Loo EPi, Yasuda S. 2018. Pattern recognition receptors and signaling in plant–microbe interactions. *The Plant Journal* **93**: 592-613.

Santiago J, Brandt B, Wildhagen M, Hohmann U, Hothorn LA, Butenko MA, Hothorn M. 2016. Mechanistic insight into a peptide hormone signaling complex mediating floral organ abscission. *Elife* **5**: e15075.
Savojardo C, Martelli PL, Fariselli P, Profiti G, Casadio R. 2018. BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic acids research* **46**: W459-W466.
Scheuring D, Viotti C, Krüger F, Künzl F, Sturm S, Bubeck J, Hillmer S, Frigerio L, Robinson DG, Pimpl P. 2011. Multivesicular bodies mature from the trans-Golgi network/early endosome in *Arabidopsis*. *The Plant Cell* **23**: 3463-3481.
Segonzac C, Monaghan J. 2019. Modulation of plant innate immune signaling by small peptides. *Current Opinion in Plant Biology* **51**: 22-28.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**: 2498-2504.
Sharma A, Hussain A, Mun B-G, Imran QM, Falak N, Lee S-U, Kim JY, Hong JK, Loake GJ, Ali A. 2016. Comprehensive analysis of plant rapid alkalization factor (RALF) genes. *Plant Physiology and Biochemistry* **106**: 82-90.
Shikata M, Matsuda Y, Ando K, Nishii A, Takemura M, Yokota A, Kohchi T. 2004. Characterization of Arabidopsis ZIM, a member of a novel plant‐specific GATA factor gene family. *Journal of experimental botany* **55**: 631-639.
Shinano T, Komatsu S, Yoshimura T, Tokutake S, Kong F-J, Watanabe T, Wasaki J, Osaki M. 2011. Proteomic analysis of secreted proteins from aseptically grown rice. *Phytochemistry* **72**: 312-320.
Shinohara H, Matsubayashi Y. 2013. Chemical synthesis of *Arabidopsis* CLV3 glycopeptide reveals the impact of hydroxyproline arabinosylation on peptide conformation and activity. *Plant and Cell Physiology* **54**: 369-374.
Smith SE, Read DJ. 2010. *Mycorrhizal symbiosis*. Academic press.
Sperschneider J, Dodds PN, Singh KB, Taylor JM. 2018. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytologist* **217**: 1764-1778.
Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y. 2007. How many genes are there in plants (... and why are they there)? *Current Opinion in Plant Biology* **10**: 199-203.
Stergiopoulos I, Wit PJGMd. 2009. Fungal effector proteins. *Annual Review of Phytopathology* **47**: 233-263.
Tabata R, Sawa S. 2014. Maturation processes and structures of small secreted peptides in plants. *Frontiers in Plant Science* **5**: 311.
Takahashi F, Suzuki T, Osakabe Y, Betsuyaku S, Kondo Y, Dohmae N, Fukuda H, Yamaguchi-Shinozaki K, Shinozaki K. 2018. A small peptide modulates stomatal control via abscisic acid in long-distance signalling. *Nature* **556**: 235-238.
Tavormina P, De Coninck B, Nikonorova N, De Smet I, Cammue BPA. 2015. The plant peptidome: An expanding repertoire of structural features and biological functions. *The Plant Cell* **27**: 2095-2118.

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M. 2004. MAPMAN: a user‐driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* **37**: 914-939.
Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. 2020. Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **18**: 607-621.
Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A. 2015. Tissue-based map of the human proteome. *Science* **347**.
van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT et al. 2014. Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**: 6589-6631.
Vayssières A, Pěnčík A, Felten J, Kohler A, Ljung K, Martin F, Legué V. 2015. Development of the poplar-Laccaria bicolor ectomycorrhiza modifies root auxin metabolism, signaling, and response. *Plant physiology* **169**: 890-902.
Viklund H, Bernsel A, Skwark M, Elofsson A. 2008. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**: 2928-2929.
Villalobos Solis MI, Poudel S, Bonnot C, Shrestha HK, Hettich RL, Veneault-Fourrey C, Martin F, Abraham PE. 2020. A viable new strategy for the discovery of peptide proteolytic cleavage products in plant-microbe interactions. *Molecular Plant-Microbe Interactions* **33**: 1177-1188.
Viotti C, Krüger F, Krebs M, Neubert C, Fink F, Lupanga U, Scheuring D, Boutté Y, Frescatada-Rosa M, Wolfenstetter S. 2013. The endoplasmic reticulum is the main membrane source for biogenesis of the lytic vacuole in *Arabidopsis*. *The Plant Cell* **25**: 3434-3449.
Wang B, Qiu Y-L. 2006. Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza* **16**: 299-363.
Wang D, Griffitts J, Starker C, Fedorova E, Limpens E, Ivanov S, Bisseling T, Long S. 2010a. A nodule-specific protein secretory pathway required for nitrogen-fixing symbiosis. *Science* **327**: 1126-1129.
Wang H, Zhuang X, Wang X, Law AHY, Zhao T, Du S, Loy MM, Jiang L. 2016. A distinct pathway for polar exocytosis in plant cell wall formation. *Plant Physiology* **172**: 1003-1018.
Wang J, Ding Y, Wang J, Hillmer S, Miao Y, Lo SW, Wang X, Robinson DG, Jiang L. 2010b. EXPO, an exocyst-positive organelle distinct from multivesicular endosomes and autophagosomes, mediates cytosol to cell wall exocytosis in *Arabidopsis* and tobacco cells. *The Plant Cell* **22**: 4009-4030.
Wang L, Einig E, Almeida-Trapp M, Albert M, Fliegmann J, Mithöfer A, Kalbacher H, Felix G. 2018a. The systemin receptor SYR1 enhances resistance of tomato against herbivorous insects. *Nature Plants* **4**: 152-156.
Wang P, Yao S, Kosami Ki, Guo T, Li J, Zhang Y, Fukao Y, Kaneko‐Kawano T, Zhang H, She YM. 2020. Identification of endogenous small peptides involved in rice

immunity through transcriptomics- and proteomics-based screening. *Plant Biotechnology Journal* **18**: 415-428.
Wang X, Chung KP, Lin W, Jiang L. 2018b. Protein secretion in plants: conventional and unconventional pathways and new techniques. *Journal of Experimental Botany* **69**: 21-37.
Wang YH, Irving HR. 2011. Developing a model of plant hormone interactions. *Plant Signaling & Behavior* **6**: 494-500.
Weerawanich K, Webster G, Ma JK, Phoolcharoen W, Sirikantaramas S. 2018. Gene expression analysis, subcellular localization, and in planta antimicrobial activity of rice (Oryza sativa L.) defensin 7 and 8. *Plant Physiology and Biochemistry* **124**: 160-166.
Whitewoods C. 2021. Evolution of CLE peptide signalling. *Seminars in Cell & Developmental Biology* **109**: 12-19.
Whitford R, Fernandez A, De Groodt R, Ortega E, Hilson P. 2008. Plant CLE peptides from two distinct functional classes synergistically induce division of vascular cells. *Proceedings of the National Academy of Sciences* **105**: 18625-18630.
Whitford R, Fernandez A, Tejos R, Pérez AC, Kleine-Vehn J, Vanneste S, Drozdzecki A, Leitner J, Abas L, Aerts M. 2012. GOLVEN secretory peptides regulate auxin carrier turnover during plant gravitropic responses. *Developmental Cell* **22**: 678-685.
Wilson BA, Thornburg CC, Henrich CJ, Grkovic T, O'Keefe BR. 2020. Creating and screening natural product libraries. *Natural Product Reports* **37**: 893-918.
Xue L-J, Alabady MS, Mohebbi M, Tsai C-J. 2015. Exploiting genome variation to improve next-generation sequencing data analysis and genome editing efficiency in *Populus tremula× alba* 717-1B4. *Tree Genetics & Genomes* **11**: 1-8.
Yang G, Liu N, Lu W, Wang S, Kan H, Zhang Y, Xu L, Chen Y. 2014. The interaction between arbuscular mycorrhizal fungi and soil phosphorus availability influences plant community productivity and ecosystem stability. *Journal of Ecology* **102**: 1072-1082.
Yang X, Kalluri UC, DiFazio SP, Wullschleger SD, Tschaplinski TJ, Cheng MZ-M, Tuskan GA. 2009. Poplar genomics: state of the science. *Critical Reviews in Plant Science* **28**: 285-308.
Yang X, Medford JI, Markel K, Shih PM, De Paoli HC, Trinh CT, McCormick AJ, Ployet R, Hussey SG, Myburg AA et al. 2020. Plant biosystems design research roadmap 1.0. *BioDesign Research* **2020**: 8051764.
Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, Adams RM, Shah MB, Hettich RL, Lindquist E et al. 2011. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Research* **21**: 634-641.
Yu CS, Chen YC, Lu CH, Hwang JK. 2006. Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics* **64**: 643-651.
Zhang H, Zhang L, Gao B, Fan H, Jin J, Botella MA, Jiang L, Lin J. 2011. Golgi apparatus-localized synaptotagmin 2 is required for unconventional secretion in *Arabidopsis*. *PLoS One* **6**: e26477.

Zhang L, Ni H, Du X, Wang S, Ma XW, Nürnberger T, Guo HS, Hua C. 2017. The Verticillium-specific protein VdSCP7 localizes to the plant nucleus and modulates immunity to fungal infections. *New Phytologist* **215**: 368-381.
Zhang L, Xing J, Lin J. 2019a. At the intersection of exocytosis and endocytosis in plants. *New Phytologist* **224**: 1479-1489.
Zhang Y, Jia C, Fullwood MJ, Kwoh CK. 2020. DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. *Briefings in Bioinformatics* doi:10.1093/bib/bbaa039.
Zhang Y, Malzahn AA, Sretenovic S, Qi Y. 2019b. The emerging and uncultivated potential of CRISPR technology in plant science. *Nature Plants* **5**: 778-794.
Zhang Y, Qi Y. 2020. Diverse systems for efficient sequence insertion and replacement in precise plant genome editing. *BioDesign Research* **2020**: 8659064.
Zhao L, Poschmann G, Waldera-Lupa D, Rafiee N, Kollmann M, Stühler K. 2019. OutCyte: a novel tool for predicting unconventional protein secretion. *Scientific Reports* **9**: 19448.
Zhou B, Benbow HR, Brennan CJ, Arunachalam C, Karki SJ, Mullins E, Feechan A, Burke JI, Doohan FM. 2020. Wheat encodes small, secreted proteins that contribute to resistance to Septoria tritici blotch. *Frontiers in Genetics* **11**: 469.
Zhou K. 2019. Glycosylphosphatidylinositol-anchored proteins in Arabidopsis and one of their common roles in signaling transduction. *Frontiers in plant science* **10**: 1022.
Zhou P, Silverstein KA, Gao L, Walton JD, Nallu S, Guhlin J, Young ND. 2013. Detecting small plant peptides using SPADA (small peptide alignment discovery application. *BMC Bioinformatics* **14**: 335.
Zhu M, Gribskov M. 2019. MiPepid: MicroPeptide identification tool using machine learning. *BMC bioinformatics* **20**: 1-11.
Zhu Y, Song D, Zhang R, Luo L, Cao S, Huang C, Sun J, Gui J, Li L. 2020. A xylem-produced peptide PtrCLE20 inhibits vascular cambium activity in *Populus*. *Plant Biotechnology Journal* **18**: 195-206.
Ziemann S, van der Linde K, Lahrmann U, Acar B, Kaschani F, Colby T, Kaiser M, Ding Y, Schmelz E, Huffaker A. 2018. An apoplastic peptide activates salicylic acid signalling in maize. *Nature Plants* **4**: 172-180.

APPENDIX

Table A1. A list of representative small secreted proteins that have been experimentally confirmed in plants.

Plant species	Protein name	Gene locus	Category	Gene family (Pfam ID)	Reference
Arabidopsis thaliana	CEP1	AT1G47485	91		(Ohyama et al. 2008b)
Arabidopsis thaliana	CLV3	AT2G27250	96	PF11250	(Kondo et al. 2006)
Arabidopsis thaliana	EPLF9	AT4G12970	102	PF16851	(Hunt et al. (2010)
Arabidopsis thaliana	EPF1	AT2G20875	104	PF13912	(Hara et al. 2007)
Arabidopsis thaliana	GLV6	AT2G03830	123		(Fernandez et al. 2015)
Arabidopsis thaliana	LTP1	AT2G38540	118	PF00234	(Potocka et al. 2012)
Arabidopsis thaliana	PREPIP1	AT4G28460	72		(Hou et al. 2014)
Arabidopsis thaliana	PREPIP2	AT4G37290	84		(Hou et al. 2014)
Arabidopsis thaliana	PROPEP1	AT5G64900	92	PF00879	(Huffaker et al. 2006)
Arabidopsis thaliana	PROPEP2	AT5G64890	109	PF00879	(Ross et al. 2014)
Arabidopsis thaliana	PROPEP3	AT5G64905	96	PF00879	(Nakaminami et al. 2018)
Arabidopsis thaliana	PSK1	AT1G13590	87	PF06404	(Mosher et al. 2013)
Arabidopsis thaliana	RALF1	AT1G02900	120	PF05498	(Sharma et al. 2016)
Arabidopsis thaliana	RGF1	AT5G60810	116		(Matsuzaki et al. 2010)
Arabidopsis thaliana	IDA1	AT3G25655	86		(Santiago et al. 2016)
Medicago truncatula	NCR169	Medtr7g029760	61	PF07127	(Horváth et al. 2015)
Oryza sativa	DEF7	LOC_Os02g41904.1	80	PF00304	(Weerawanich et al. 2018)
Populus trichocarpa	CLE20	Potri.014G156600	74		(Zhu et al. 2020)
Solanum lycopersicum	CAPE1	Solyc00g174340	159	PF00188	(Chen et al. 2014)
Zea mays	PROZIP1	AC210027.3_FG003	137		(Ziemann et al. 2018)

Table A2. A list of representative computational resources and tools for predicting plant SSPs.

Type	Name	Description	Website	Reference
Database	MtSSPdb	SSPs in *Medicago truncatula*	https://mtsspdb.noble.org	(Boschiero et al. 2020)
Database	PlantSecKB	All secreted proteins in multiple species	http://proteomics.ysu.edu/secretomes/plant.php	(Lum et al. 2014)
Database	OrysPSSP	SSPs in *Oryza sativa*	http://www.genoportal.org/PSSP/index.do	(Pan et al. 2012)
Standalone Package	DeepCPP	Predicting RNA coding potential	https://github.com/yuuuuzhang/DeepCPP	(Zhang et al. 2020)
Online tool	SignalP-5.0	Predicting signal peptides	http://www.cbs.dtu.dk/services/SignalP/	(Almagro Armenteros et al. 2019)
Online tool	SecretomeP	Predicting non-classical protein secretion	http://www.cbs.dtu.dk/services/SecretomeP/	(Nielsen et al. 2019)
Online tool	OutCyte	Predicting unconventional protein secretion	http://www.outcyte.com/	(Zhao et al. 2019)
Online tool	ApoplastP	Predicting effectors and plant proteins in the apoplast using machine learning	http://apoplastp.csiro.au	(Sperschneider et al. 2018)
Online tool	DeepLoc	Prediction of subcellular localization of eukaryotic proteins	http://www.cbs.dtu.dk/services/DeepLoc/	(Almagro Armenteros et al. 2017)
Online tool	LocTree3	Predicting subcellular localizations	https://rostlab.org/services/loctree3/	(Goldberg et al. 2014)
Online tool	BUSCA	Predicting subcellular localizations	http://busca.biocomp.unibo.it	(Savojardo et al. 2018)
Online tool	Plant-mSubP	Predicting subcellular localizations	http://bioinfo.usu.edu/Plant-mSubP/	(Sahu et al. 2020)
Standalone package	MEMSAT-SVM	Transmembrane helix topology	https://github.com/psipred/MemSatSVM	(Nugent and Jones 2012)

Table A2. continued

		prediction; identifying the cytosolic and extra-cellular loops.		

VITA

Xiaoli Hu was born in Panzhihua, the southern city of Sichuan province, China. He earned a Bachelor of Agronomy degree at Sichuan Agricultural University after four years study in the Agricultural Resource and Environment program. In fall of 2015, he qualified for an exemption of Master's Entrance Exam and enrolled in China Agricultural University to begin graduate study. He worked in Dr. Zhenhai Han's lab and focused on learning how different combinations of scion-stock of apples impact root structure of tree. In 2017, he graduated from Department of Horticulture at China Agriculture University and decided to pursue his dream in the USA. In the same year, he began his Ph.D. studies at the University of Tennessee-Knoxville. He was co-advised by Dr. Zong-Ming (Max) Cheng and Dr. Xiaohan Yang to identify and investigate plant small secreted proteins which are involved in plant-fungi symbiosis.